People detection methods for intelligent multi-Camera surveillance systems

Muhammad Owais Mehmood

ECOLE CENTRALE DE LILLE

THESE

présentée en vue

d'obtenir le grade de

DOCTEUR

en

Spécialité : Automatique, Génie Informatique, Traitement du Signal et Images

par

Muhammad Owais Mehmood

DOCTORAT DELIVRE PAR L'ECOLE CENTRALE DE LILLE

Titre de la thèse :

Detection de personnes pour des systèmes de videosurveillance multi-caméra intelligents

Soutenue le 28 Septembre 2015 devant le jury d'examen :

Président	Samia, Bouchafa-Bruneau, Professeur, Université d'Evry Val d'Essonne
Rapporteur	Roland, Chapuis, Professeur, Université Blaise Pascal, Aubiere
Rapporteur	François, Bremond, Directeur de Recherche, INRIA, Sophia Antipolis
Membre	Stéphane, Lecoeuche, Professeur, Institut Mines-Télécom, Douai
Membre	Patrick, Sayd, Ingénieur de recherche, CEA List, Palaiseau
Invité	Mounim A., El Yacoubi, Directeur d'études, Telecom SudParis, Evry
Directeur de thèse	Pierre, Chainais, Maître de Conférences, Ecole Centrale Lille
Co-Directeur de thèse	Catherine, Achard, Maître de Conférences, UPMC, Paris
Encadrant	Sébastien, Ambellouis, Chargé de recherche, IFSTTAR, Villeneuve d'Ascq

Thèse préparée dans le Laboratoire Électronique Ondes et Signaux pour les Transports (LEOST) de l'Institut français des sciences et technologies des transports, de l'aménagement et des réseaux (IFSTTAR)

Ecole Doctorale SPI 072

PRES Université Lille Nord-de-France

ECOLE CENTRALE DE LILLE

THESIS

Submitted for the degree of

DOCTOR

in

Domain : Automatic Control, Computer Science, Signal and Image Processing

by

Muhammad Owais Mehmood

AWARDED BY ECOLE CENTRALE DE LILLE

TITLE :

People Detection Methods For Intelligent Multi-Camera Surveillance Systems

Defended on 28 September 2015 in the presence of the jury :

President	Samia, Bouchafa-Bruneau, Professor, Université d'Evry Val d'Essonne
Reviewer	Roland, Chapuis, Professeur, Université Blaise Pascal, Aubiere
Reviewer	François, Bremond, Director of Research, INRIA, Sophia Antipolis
Member	Stéphane, Lecoeuche, Professor, Institut Mines-Télécom, Douai
Member	Patrick, Sayd, Research Engineer, CEA List, Palaiseau
Invited	Mounim A., El Yacoubi, Director of Studies, Telecom SudParis, Evry
Thesis Director	Pierre, Chainais, Associate Professor, Ecole Centrale Lille
Thesis Co-Director	Catherine, Achard, Associate Professor, UPMC, Paris
Thesis Supervisor	Sébastien, Ambellouis, Senior Researcher, IFSTTAR, Villeneuve d'Ascq

Thesis prepared at the Laboratory of Electronics, Waves and Signal Processing for Transport (LEOST), French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR)

Doctoral School SPI 072

PRES Université Lille Nord-de-France

"From above it is not bright;
From below it is not dark:
An unbroken thread beyond description.
It returns to nothingness.
The form of the formless,
The image of the imageless,
It is called indefinable and beyond imagination."

Lao-Tzu

Résumé

La détection de personnes dans les vidéos est un défi bien connu du domaine de la vision par ordinateur avec un grand nombre d'applications telles que le développement de systèmes de surveillance visuels. Même si les détecteurs monoculaires sont plus simples à mettre en place, ils sont dans l'incapacité de gérer des scènes complexes avec des occultations, une grande densité de personnes ou des scènes avec beaucoup de profondeur de champ menant à une grande variabilité dans la taille des personnes. Dans cette thèse, nous étudions la détection de personnes par un système multicaméras et plus particulièrement, l'utilisation de cartes d'occupation probabilistes créées en fusionnant les différentes vues grâce à la connaissance de la géométrie du système. La détection à partir de ces cartes d'occupation amène cependant de fausses détections dues aux différentes projections. Celles-ci, bien connues dans la littérature, sont dénommées "fantôme". Aussi, nous proposons deux nouvelles techniques remédiant à ce problème et améliorant la détection des personnes. La première utilise une déconvolution par un noyau dont la forme varie spatialement tandis que la seconde est basée sur un principe de validation d'hypothèse. Ces deux approches n'utilisent volontairement pas l'information temporelle qui pourra être ré-introduite par la suite dans des algorithmes de suivi. Les deux approches ont été validées dans des conditions difficiles présentant des occultations, des encombrements plus ou moins denses et de fortes variations dans les réponses colorimétriques des caméras. Une comparaison avec d'autres méthodes de l'état de l'art a également été menée sur trois bases de données publiques, validant les méthodes proposées dans le cadre des transports en commun, à savoir, la surveillance d'une gare et d'un aéroport.

Mots-clefs: Géométrie multi-vues, Fusion de capteurs, Reconnaissance des Formes, Détection d'objects, Surveillance.

Abstract

People detection is a well-studied open challenge in the field of Computer Vision with various applications such as in the visual surveillance systems. Monocular detectors have limited ability to handle complexities such as occlusion, clutter, scale, density. Ubiquitous presence of cameras and computational resources fuel the development of multi-camera detection systems. In this thesis, we study the multi-camera people detection; specifically, the use of multi-view probabilistic occupancy maps based on the camera calibration. Occupancy maps allow multi-view geometric fusion of several camera views. Detection with such maps produces several false detections and we study this phenomenon: ghost pruning. To this end, we propose two novel techniques in order to improve multi-view detection based on: (a) kernel deconvolution, and (b) occupancy shape modeling. We perform non-temporal, multi-view reasoning in occupancy maps to recover accurate positions of people in challenging conditions such as occlusion, clutter, lighting, and camera variations. We show improvements in people detections across three challenging datasets for visual surveillance including comparison with state-of-the-art techniques. We show the application of this work in exigent transportation scenarios i.e. people detection for surveillance at train stations and airports.

Keywords: Multi-view Geometry, Sensor Fusion, Pattern Recognition, Object Detection, Surveillance.

Acknowledgements

First and foremost, I would like to express my thanks to the thesis supervisors: Prof. Pierre Chainais, Prof. Catherine Achard, and Dr. Sébastien Ambellouis. This thesis could not have been possible without their support, guidance, assistance, management, patience, dedication, and interest. Prof. Pierre Chainais has been kind and supportive as we moved the thesis to École Centrale de Lille. I can not fail to mention the initiatives, research ideas, prompt yet efficient reviews, handling my last minute requests with kindness, and the enthusiasm exhibited for this thesis by Prof. Catherine Achard. She must be praised even more as she managed most of this from Paris. Finally, Dr. Sébastien Ambellouis who has been a tremendous support, guide, even an active collaborator, and also a friend. He has shown me incredible kindness in managing affairs ranging from research, programming, this thesis, the administrative affairs, and also teaching, advising me about the various non-technical skills.

Secondly, I wish to thank the institute for granting me a fellowship to complete this thesis. Moreover, I also appreciate the contributions of the DéGIV project towards my work.

I also wish to thank my thesis reviewers, Prof. Roland Chapuis, and Dr. François Bremond for agreeing to review my thesis. I am also grateful to the other members of the thesis committee: Prof. Stéphane Lecoeuche, Prof. Samia Bouchafa-Bruneau, Dr. Patrick Sayd, and Dr. Mounim A. El Yacoubi for their willingness to make time out of their schedules for my work.

I am grateful to the colleagues at the office. They made my journey of this Ph.D. a whole lot easier. Hassanein was there as a peer, even if from another field, to share non-technical ideas such as writing, LaTeX, and making pretty curves. A special thanks goes to Adil for helping me manage many, many, many administrative affairs of this thesis, making lunches interesting even when I was writing this thesis, and also for his insightful discussions.

It was in 2008 when Dr. Zeeshan Zia introduced me to the field of computer vision. I know him for 9 years now and he has been the best mentor in my life. He has kept me motivated for research, and for vision research, and has had a major impact on my life. Similarly, since 2011, Dr. Rahat Khan has been there guiding me, motivating me, not only in vision, but, also life in Europe, and France.

I must also thank my French language teacher Madam Véronique Attagnant. It was only because of her personal attention that I finally broke my shyness with the language. She explained to me that I must keep moving on, keep learning, remain inspired, even if the path is not always smooth. I wish I could have time and opportunity to attend more of her courses.

Thanks to many, many of my friends at Lille. I will not mention names out of the fear of missing someone out. I actually had the honor to meet, discuss, learn from each and everyone of you. If you're reading this, are not my colleague, and we met in Lille, be sure, you have been acknowledged. However, I am bound to mention Pierre. I remember he helped me in November 2012 when I might have been forced to discontinue my studies due to unforeseen personal issues. Thanks to him, it never happened and since then he has been a social mentor to me. He has contributed to this thesis with each and every non-technical skill possible; and, has transformed me into a much, much better person. I must also thank the people who agreed to proof-read this thesis at a very short notice: Pierre, Adil, Rahat, Hassanein, and Zeeshan.

Last but not the least, how can I forget to mention my three sisters and parents. I am what I am because of my parents and they have done unimaginable sacrifices, the gist of which can never be put in words or actions, so, thank you so much *Maa* and Dad!!

Contents

Résumé ii

Abstract iii

Acknowledgements iv

Contents vi

List of Figures ix

List of Tables xi

List of Algorithms xii

1 Introduction 1
- 1.1 Background and Motivation . . . 1
- 1.2 Objectives . . . 3
- 1.3 Applications . . . 5
- 1.4 Thesis Outline . . . 6
- 1.5 Contributions . . . 7

2 Related Work 9
- 2.1 Introduction . . . 9
- 2.2 Monocular People Detection . . . 9
- 2.3 Intelligent Multi-camera Surveillance . . . 11
  - 2.3.1 Calibration . . . 13
  - 2.3.2 Network Topology . . . 15
- 2.4 Multi-Camera People Detection . . . 16
- 2.5 Summary . . . 20

3 Ghosts in the Multi-camera Occupancy Maps 24
- 3.1 Introduction . . . 24
- 3.2 Multi-planar Projections and their combination . . . 25
  - 3.2.1 Foreground Masks . . . 25

3.2.2 Multi-planar Projections . . . 28
3.2.3 Homographic Occupancy Constraint . . . 31
3.2.4 Multi-camera Occupancy Maps and Ghosts effect . . . 33
3.3 Summary . . . 37

4 Multi-camera Occupancy Map Deconvolution 39
4.1 Introduction . . . 39
4.2 Modeling of the occupancy map . . . 40
4.3 People Detection . . . 43
4.3.1 Deconvolution . . . 43
4.3.2 Maxima Selection using Watershed . . . 46
4.4 Experimental Results . . . 48
4.4.1 Dataset . . . 49
4.4.2 Evaluation Measures . . . 50
4.4.3 Results . . . 56
4.5 Summary . . . 61

5 Ghost Pruning in the Multi-camera Occupancy Maps 64
5.1 Introduction . . . 64
5.2 Keypoint Extraction in the Occupancy Map . . . 65
5.3 Ghost Pruning with Shape Analysis . . . 66
5.4 Detection based on thresholding . . . 70
5.5 Results on PETS 2009 . . . 72
5.5.1 Study of the similarity measure . . . 72
5.5.2 Influence of the parameter τ . . . 73
5.5.3 Influence of the methods used to select the optimal threshold 73
5.5.4 Comparison with other works . . . 74
5.6 Summary . . . 79

6 Experiments on Other Datasets 81
6.1 Introduction . . . 81
6.2 Datasets . . . 81
6.2.1 PETS 2006 . . . 82
6.2.2 PETS 2007 . . . 83
6.3 Experimental Setup . . . 84
6.4 Results . . . 88
6.4.1 PETS 2006 . . . 88
6.4.2 PETS 2007 . . . 89
6.5 Summary . . . 92

7 Conclusions and Future Work 96
7.1 Summary of Key Contributions . . . 96
7.2 Limitations of the Work . . . 98
7.3 Future Work . . . 99

A Résumé en francais **102**

Bibliography **109**

List of Figures

1.1 Example of an abnormal situation at a train station 2
1.2 Example of challenges faced in intelligent surveillance 3
1.3 Example of a commercial multi-camera surveillance system 4

2.1 Example of datasets used for monocular people detection 10
2.2 Some algorithms employed for automated multi-camera surveillance 12
2.3 Overlap in camera field of view 14
2.4 Fusion in multi-camera networks 15

3.1 Block diagram for detection using occupancy maps 25
3.2 Local binary patterns 26
3.3 Example of a PETS 2009 frame and the foreground masks 29
3.4 Camera projections at different planes of the scene 30
3.5 Camera projections at different planes of the scene 31
3.6 Example of multi-planar projections 31
3.7 Proposition for the occupancy maps 32
3.8 Homographic occupancy constraint 33
3.9 Combination of planar projections 34
3.10 Ghosting phenomenon in an arbitrary scenario 35
3.11 Ghosting phenomenon in different planes 36
3.12 Illustration of an occupancy map with ghosts 37

4.1 Block diagram for primitive detection process applied on the occupancy maps 40
4.2 Illustration of the primitive based detection process on the multi-camera occupancy map 41
4.4 Example images during various stages of the primitive detection process 45
4.5 Illustration of the false detection 47
4.6 Keypoint extraction process 47
4.7 Location of PETS 2009 cameras on Google Maps 49
4.8 Illustration of calibration error in PETS 2009 50
4.9 Frame 683 from PETS 2009 dataset 51
4.10 Frame 752 from PETS 2009 dataset 52
4.11 Illustration of the false detection 54
4.12 Frame 593 from PETS 2009 dataset 57

4.13 Frame 683 from PETS 2009 dataset . . . 58
4.14 Frame 752 from PETS 2009 dataset . . . 59
4.15 Evaluation of the MOD method with different parameter settings on PETS 2009 . . . 60
4.16 Examples of the blurring effect . . . 63

5.1 Block diagram for detection process by shaping the occupancy maps across the keypoints. Here synergy shape model refers to the proposed synthetic shapes. . . . 65
5.2 Concept of the Occupancy Shape Model as visualized with the occupancy maps . . . 67
5.4 Example of Occupancy Shape Model of ghosts and pedestrians . . . 69
5.5 Taxonomy of data clustering techniques . . . 71
5.6 Evaluation of the proposed Occupancy Shape Model with different parameter settings on PETS 2009 . . . 75
5.7 Results obtained on frame 593 of PETS 2009 dataset via OSM method 76
5.8 Results obtained on frame 752 of PETS 2009 dataset via OSM method 77
5.9 Results obtained on frame 784 of PETS 2009 dataset via OSM method 78
5.10 Results obtained on frame 683 of PETS 2009 dataset via OSM method 80

6.1 Sequence selection for PETS 2006 . . . 82
6.2 Area of Interest selection for PETS 2006 . . . 84
6.3 Example of a frame in PETS 2006. . . . 85
6.4 Example of foreground masks on PETS 2006 . . . 85
6.5 Sequence selection for PETS 2007 . . . 86
6.6 Area of Interest selection for PETS 2007 . . . 86
6.7 Example of a frame in PETS 2007. . . . 87
6.8 Example of foreground masks on PETS 2007 . . . 87
6.9 Evaluation of the MOD method with different parameter settings on PETS 2006 . . . 88
6.10 Comparison between MOD and OSM for PETS 2006 . . . 90
6.11 Examples of the estimated pedestrian locations in the PETS 2006 dataset (MOD) . . . 91
6.12 Evaluation of the MOD method with different parameter settings on PETS 2007 . . . 92
6.13 Examples of the estimated pedestrian locations in the PETS 2006 and PETS 2007 datasets (OSM) . . . 94
6.14 Example of the estimated pedestrian locations in the PETS 2007 datasets (MOD) . . . 95

7.1 Vehicle Detection with 3D Cuboids . . . 99

List of Tables

2.1 Summary of some monocular systems discussed in the related work. 22
2.2 Summary of some multi-camera systems discussed in the related work. 23

4.2 Results of the MOD method on PETS 2009 dataset 61

5.1 Comparison of the different similarity measures 72
5.2 Comparison of proposed OSM with other techniques 73
5.3 Comparison of the different clustering techniques 74

6.1 Results on the PETS 2006 dataset 88
6.2 Results on the PETS 2007 dataset 89
6.3 Summary of the results 93

A.2 Results of the MOD method on PETS 2009 dataset 106
A.3 Comparison of proposed OSM with other techniques 107

List of Algorithms

1 Watershed based maxima selection 46
2 People detection by deconvolution 48
3 People detection by ghost pruning in multi-camera occupancy maps 71

Chapter 1

Introduction

1.1 Background and Motivation

Intelligent automated systems, such as for visual surveillance, is an active area of interest for human societies with the web of attention spiraling in public, military, commercial and research circles [1]. Ideally, the goals of automated surveillance systems are to perform detection, tracking, classification, and recognition of the objects in the scene. Such goals lead to the endowment towards higher-level semantic tasks such as human behavior analysis and activity recognition. Applications of visual surveillance range from crime prevention, traffic control [2] to monitoring patients at hospital [3], and the children at home [4]. Multiple domains of research contribute to the field of intelligent surveillance. These domains include computer vision, image processing, pattern recognition, artificial intelligence, (big) data management, signal processing, telecommunications, embedded systems, sensor design and electronics, and socio-ethical studies.

Vision is one of the five human senses and an important one for large scale usage; it is often said: *a picture is worth a thousand words*. Computer Vision had been thought of as an easy problem of artificial intelligence in the early sixties. The slow initial pace of progress in vision was well explained by the Moravec's paradox arguing that tasks requiring high-level reasoning need enormous computational resources. Nowadays, computer vision has emerged as a discipline itself, trying to tackle the tasks of automated vision. Despite the challenges, some of which remain an open area of research even today, substantial progress has been made.

Research on camera networks, like visual surveillance in general or people detection in particular, has received much attention in computer vision literature [5]. For example, works on multi-view camera geometry [6] remain a holy grail of overlapping multi-sensor visual processing [7].

People detection is an integral element of any machine's environment and also plays a key part in the surveillance. The numbers of fatalities in France at the level crossing, or the number of pedestrian fatalities in US related to traffic crash could be a possible indicator to how we can deploy such people detectors inside our cars or at transportation infrastructure. The development of the video surveillance inside the trains is another context where it is required to propose such automatic video functions to improve the security and the comfort pf passengers. We may consider the developments of these tasks of paramount importance, requiring high accuracies, through one camera. However, such a system will have a limited field of view and resources. Recent technological developments have produced a boom in the ubiquitous installation and usage of cameras. This has been complemented by phenomenal breakthroughs in the cheap yet efficient computational resources, including the embedded processing units. Thus, the multi-camera video surveillance systems with people detection abilities have gained the interest of various segments of the community including those of the computer vision and pattern recognition researchers.

This thesis focuses on the areas of computer vision and pattern recognition in order to answer: can we further add to the performance of existing multi-view people detectors? For this purpose, we have chosen the context of visual surveillance. This thesis has been completed at the *French Institute of Science and Technology for Transport, Development and Networks*. The institute specializes in the research of transportation systems across various domains including the domain of artificial

Figure 1.1: Example of application in transportation scenario where people detection can help to identify a concernful situation. Reproduced from [8].

intelligence and vision for achieving intelligent transportation systems. Figure 1.1 shows an abnormal situation of interest for manual or automated detectors, or analysers at a train station. Therefore, one of the motivations of this thesis is to show the application of multi-view people detection in the transportation context. This work could also act as a precursor to other areas of active vision research at the various laboratories of the institute such as person identification and re-identification (LEOST-IFSTTAR), human activity recognition (LEOST), human perceptual analysis and cognition in the context of car drivers (LESCOT), and pedestrian behaviour modeling (LEPSIS).

1.2 Objectives

The aim of our work is to perform people detection across multiple camera sensors. These surveillance cameras are stationary, wide baseline, viewing planar scenes, and provide calibrated images. Such synchronised setups are popular in the visual surveillance scenarios which is a goal of this work. The following is a more fine-grained list of our aims and objectives:

- To improve multi-camera people detection using a combination of multi-view geometry and pattern recognition techniques. The problem formulation could be considered as a multi-sensor data fusion [10] or that of image

Figure 1.2: Example of challenges faced in intelligent surveillance. Reproduced from [9].

registration in the field of computer vision [11]. However, the results must show robustness to adverse and challenging conditions such as camera variations like resolution, perspectives, image quality, color variations, lighting variations, motion sensitivity, clutter, noise, occlusions, and issues arising due to density of the objects such as people of varying sizes present in the scene. Figure 1.2 illustrates some of these challenges.

- To improve the well-established occupancy map framework [12, 13] for multi-view detection. Homographic occupancy constraint provides an acceptable solution to many of the issues mentioned above for multi-view fusion and detection. However, such maps also create several erroneous detections known as ghosts in the literature [14, 15]. Our goal is thus to avoid these errors without considering temporal information.

- To demonstrate applications in the transportation scenario. For example, can we apply our work to the multiple surveillance cameras installed at the airports, train stations, etc?

Figure 1.3: Example of a commercial multi-camera surveillance system: eight camera surveillance setup in a Houston, Texas laundromat. Reproduced from [16].

1.3 Applications

Humans are an important element of consideration for computers and machines in their respective environments. People detection and their manipulation in the visual sensors paves way to many interesting applications many of which play a vital role in our societies. These applications, being in a constant state of innovation, are expected to take various new shapes and forms. The following is a non-exhaustive list of some of the areas where people detection and surveillance find their applications:

Transportation: There are several transportation applications e.g. security at the airports, train, metro, bus stations, sidewalk for pedestrians, monitoring abandoned objects like luggage inside trains, monitoring line-crossing, traffic signal violations, etc. One such scenario is the centralized visual surveillance in Manchester's transportation services [17].

Personal: People detection has applications in remote and mobile monitoring such as with security systems installed at apartments. It is also being deployed in automated cars e.g. automated driving in Mercedes Benz F015 concept vehicle or the Google Car. Microsoft Kinect is one innovation, combining visual and infrared sensors, which has reached the homes of many across the world with Xbox to revolutionize how people play video games [18, 19].

Commercial: Commercial institutions require applications designed to tackle issues of facility protection, operational monitoring, vandalism, employee safety, etc. One such application is seen in Figure 1.3. High-tech companies on Internet such as Google employ detection applications for improving user experiences [20].

Government: Public safety is a major concern for governments in the contemporary world [21]. People detection, visual surveillance are key elements in the installations of event video monitoring, safety apparatus at public locations such as streets, parks, hospitals, etc. Applications also exist of such algorithms in the Internet security systems. There are military interest in such technologies as well. Examples of military uses range from the automated robots, drones to the satellite surveillance.

Research: People detection algorithms [22] in visual modalities could further aid towards the development of other semantic contexts such as people tracking, identification, re-identification, human behavior analysis, activity recognition, etc. Similarly, the applications diverge into other related fields such as robotic vision, medical vision, virtual reality, etc.

1.4 Thesis Outline

The goal of this thesis is to contribute to the development of robust multi-camera people detectors. Exploiting temporal axis is known to improve detection results [13, 15, 23, 24]; nevertheless, the objective of the thesis is not to build a complete system rather improve the per-frame detection performance. To this end, we focus on a stronger per-frame detector. This detector, if used as a module in a bigger system, would also cause a more holistic dynamic scene understanding system to perform better. For hardware purposes, we have used static, calibrated, synchronized, centralized, and overlapping multi-camera networks [5] in a wide baseline configuration [25]. We emphasise on visual surveillance in transportation [26, 27] and public spaces [28]. We plan to achieve this using the concepts of multi-view geometry [6] and pattern recognition [29].

We begin by the presentation of the related works to this thesis covering various topics about multi-camera surveillance system; and, state-of-the-art on monocular and multi-camera people detection algorithms. Monocular methods have limited knowledge about the scene, and thus are limited in terms of detection accuracies [30]. This is where the multi-camera systems can be of help [7]. Multi-camera occupancy maps is a popular technique [12, 13] for people detection and tracking, but it has an important flaw that it generates false detections known as ghosts [14, 24]. In a first time, we present the generation process of the multi-camera occupancy maps. Further, we explain the homography constraint, which makes the detection through this method so useful. Finally, study of the limitations of the multi-occupancy maps i.e. the ghost phenomenon is presented. After covering these fundamental issues in Chapter 3, we begin with the presentation of the developed techniques.

Chapter 4 presents the multi-camera occupancy map deconvolution method. It begins with the presentation of the modelling of the process of occupancy map

creation as a convolution by a spatially-varying kernel. We also present an estimated deconvolution technique for kernel-based detection. This method extends the monocular technique presented in [31]. While the latter is unable to account for multi-camera fusion, we successfully present an extension with the multi-view reasoning. In chapter 5, we present a method for ghosts pruning in the multi-camera occupancy maps. For this purpose, we introduce a novel method for the robust detection of candidates in the occupancy map. We also propose the generation of a synthetic shape model in the occupancy map at these candidate locations. Finally, there is a proposition of a similarity measure and clustering on the measures distribution to perform automatic thresholding for people detection. Experimental results and analysis have been performed on the popular, public, and challenging PETS 2009 dataset [28].

For further validation of our techniques, we have also presented experimentations on two other datasets: PETS 2006 [26] and PETS 2007 [26]. Moreover, we have also performed ground truth annotation. Experimental results and analysis are obtained from both these datasets operating in various conditions, and in transportation scenarios. We show the success of our method at improving detection rates despite the variable, and challenging conditions. Finally, we also present limitations of our work, and the development of future ideas in light of such limitations. Our method is a contribution in the area of multi-view geometry based people detection. It accounts for challenging conditions and exhibits significant potential as a module for further higher-level tasks such as: tracking, identification, activity/behavior analysis, in context of visual surveillance.

1.5 Contributions

The original research contributions of this thesis are listed as follows:

- Modelling of the detection process performed on the occupancy maps as a convolution with spatially-varying kernels. This novel method performs multi-view geometric reasoning and studies the shapes of people in the occupancy maps.

- People detection based on an approximation of the deconvolution, this method allows us to avoid the ghost detection, a well known problem in the literature. False detections due to ghost decrease the per-frame detection rates. Therefore, this contribution enhances the detection efficiencies.

- Reformulation of the problem of detection as a problem of hypothesis checking. This hypothesis evaluation is limited to the key locations on the occupancy map. We have also proposed a method for detection of these key locations.

- Analysis and quantitative evaluation of the proposed methods on three challenging public datasets — PETS 2006, PETS 2007, and PETS 2009. Further, we have provided ground truth annotation of multi-sensor people detection for the sequences PETS 2006 and PETS 2007.

Chapter 2

Related Work

2.1 Introduction

In this chapter, we present a literature study in the context of our research aims and propositions. We introduce the research works and concepts regarding the monocular people detection, intelligent multi-camera surveillance methods, and multi-camera people detection. We show the current achievements and limitations of these works and introduce some concepts of the multi-camera networks to aid in the development of our research objectives. We focus on improved per-frame detection rates, without the use of any temporal information. It is also helpful for holistic visual systems to have perfect detection as input. Some of these ideas are also discussed later in the thesis.

2.2 Monocular People Detection

People detection is an active area of research in computer vision; there is an ever-growing number of approaches in the literature which try to address this issue [30, 32]. Due to the considerable variety of methodologies and datasets employed for performing people detection, the fundamental questions on how well these monocular detectors work, their standardised comparison, and identification of failure cases are difficult to answer. For answering these questions: sixteen pre-trained pedestrian detectors have been tested in a standardised manner on six popular public datasets in [30]. Example of the datasets and scenarios in

which the current pedestrian detectors work is shown in Figure 2.1. The authors have identified seven specific research directions in the field of monocular people detector; we present a summary of these research directions [30] for monocular detection below:

- **Scales.** People detection is sensitive to the scaling of the humans visible in the images. The performance decreases at lower scales. According to the authors, the detectors do not perform well for a scale between 30 and 80 pixels.

Figure 2.1: Example of datasets used for monocular people detection. The bounding boxes are obtained from the ground truth annotations. Modified and reproduced from [30].

- **Occlusion.** Monocular people detectors do not completely account for occlusion even if it is not severe.
- **Motion features.** Monocular detectors can perform better with the utilisation of motion features.
- **Temporal features.** Detection techniques do not make use of temporal features but few consecutive frames could help the detection.
- **Context.** Geometric constraints such as those introduced with the ground plane may improve the detection.
- **Novel features.** Novelties in the feature extraction domain could benefit the people detectors.
- **Datasets.** Further research is required to study the effects of the different types and amount of data used for people detectors, specially those using the machine learning techniques.

The authors in [30] have discussed monocular people detectors in single still images and shown that most of them are based on learning. In [33], the authors, propose a survey on how these people detectors have been employed in a complete system. The multi-camera systems, in comparison to single camera detectors, can inherently provide more information and thus have been exploited to perform more robust people detection [7]. For example, they have better ability to handle scales, occlusion, motion feature, temporal features, and the ground plane context [13]. Introducing tracking, as presented in [34], may alleviate the problems arising due to detection, but it does not completely solve it. Nevertheless, in this thesis, our problem formulation is to focus on per-frame detection [35] i.e. an algorithm to find bounding boxes for people on the ground plane in the 3D world, and across all the camera views, without the use of temporal features.

2.3 Intelligent Multi-camera Surveillance

This section addresses the various goals and techniques for intelligent multi-camera surveillance system. This area of research is multi-disciplinary. It involves domains such as vision, signal processing, sensor networks, and embedded systems. Therefore, it could be possible to classify and list these tasks in numerous ways. We

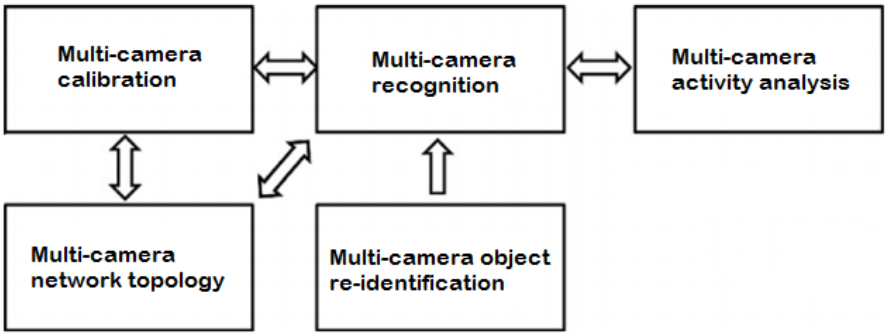


Figure 2.2: Some algorithms employed for automated multi-camera surveillance. Flow between various modules is indicated by the arrows. Modified and reproduced from [7].

use the methodology and presentation in terms of computer vision, as introduced in [7]. Figure 2.2 shows the various technologies used in the multi-camera surveillance system. We now briefly present each of them below:

- **Calibration** is the registration of the various camera views to one common coordinate system. It involves concepts of multi-view geometry [6] and is a key step for various multi-camera systems including ours. We further discuss this in Section 2.3.1.
- **Network topology** is the study of the camera networks in terms of their spatial adjacency, time transitions, and overlapping or non-overlapping fields of view. We focus on overlapping camera works during the course of this thesis. More details on network topology [36] are presented in Section 2.3.2.
- **Recognition** is the study of object detection and tracking across the camera networks. We study multi-camera detection in Section 2.4. For a general study of object tracking algorithms, the literature review in [34] can be referred to.
- **Object re-identification** is the recognition of the same person across disjoint camera views at a different time and location. Re-identification could also be understood as the matching of two image regions across different views. This matching is performed primarily using appearance-based features. Object re-identification is not discussed in this thesis but further details can be seen in [37].
- **Activity analysis** performs automated recognition of various activities occurring in the scene including understanding the human behavior. The aim

is to automatically determine the abnormal activities being monitored across the multiple cameras. We will not be discussing activity analysis or behavior recognition in this work. However, a general literature review on human activity analysis can been seen in [38].

The work proposed in this thesis is focused specifically on multi-camera recognition. It is linked to the calibration step and the topology of camera network used. We will present the details of these three problematics in the next section.

2.3.1 Calibration

Camera calibration, or geometric camera calibration, in the context of 3D vision is the process used to determine the parameters of the camera. These parameters can be intrinsic (internal camera geometry, optical characteristics) or extrinsic (relationship of the 3D camera frame and position to an arbitrary world coordinate system) [39]. Camera calibration provides information to assist in the inference of 2D image coordinates from 3D, or inferring the 3D information from the image coordinates when using several cameras. Camera calibration and pose estimation are of major interest for researchers in computer vision and it finds application in various areas, for example, surveillance, structure from motion, stereo vision, robotic simultaneous localization and mapping, etc [40].

The process of calibration finds an estimation of a model for uncalibrated cameras: position and orientation parameters which provide the extrinsic information; and, focal length, image center or the principle point, and distortion coefficients [41]. Sometimes, the internal camera parameters are already provided by the camera manufacturer. In this case, the problem transforms to that of pose estimation: that is, to recover the parameters related to the position and the orientation of the sensor in question [42].

Tsai calibration [43] is an example of the well-known calibration technique employed in varying applications related to calibration and pose estimation. It can deal with both coplanar and non-coplanar points and perform both the internal and external calibrations individually. That is, if the internal parameters of the camera are already available, then Tsai method can compute the pose estimation.

There are also methods for camera auto-calibration or camera self calibration which perform the automatic determination of the internal camera parameters from multiple uncalibrated images [44]. This technique does not require a reference calibration object and can be achieved through establishing relationship between the images formed in an environment. Finally, there are also calibration techniques for active vision systems [45]. In such techniques, a specific motion pattern is performed followed by the linear computation of the internal and external parameters, using image features and motion model. Further details on these camera calibration techniques can be studied in [44, 46, 47].

For visual surveillance, one of the most common assumption is that the objects of interest are located and move on a common ground plane. Following this assumption: camera calibration and pose estimation therefore become a vital element for multi-camera surveillance. Besides calibration, if there exists an overlap between the cameras, then, it is also possible to compute planar homography for establishing correspondence between the views [48].

Temporal synchronization also affects the performance of the camera correspondences [49]. The frames acquired from several cameras have to correspond to the same moment to be sure that the objects have not moved between the frames. Time synchronization methods include hardware methods involving distributing time stamps, detecting visible flashes; or, calculation of automated temporal shifts till achievement of optimal correspondences. A generic study of time synchronization in sensor networks can be seen in [50].

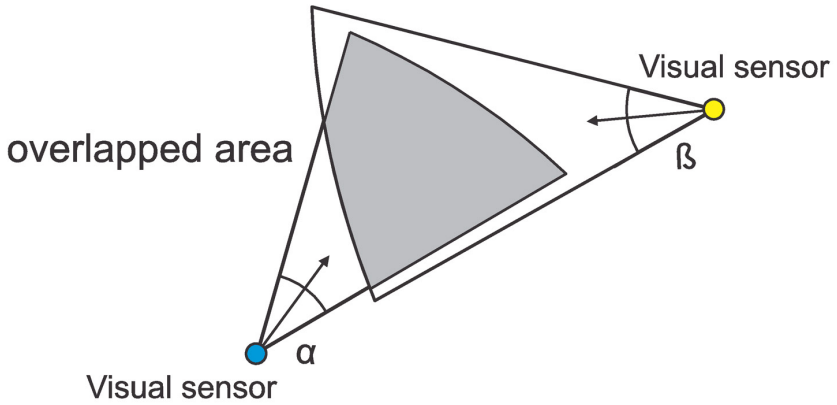


Figure 2.3: Overlapping in the camera fields of view. Reproduced from [51].

2.3.2 Network Topology

Multi-camera networks are utilised for monitoring the objects in a wide area setup or analysing objects with different view points in the scene. There are two strategies for performing the wide-area surveillance: overlapping and non-overlapping camera networks [51]. Figure 2.3 illustrates the concept of the overlapping region in the Field of View (FOV) of two cameras. Determination of this particular configuration along with the location and orientation of cameras provides an useful knowledge for automation [52]. In case of overlapping cameras, the concepts of multi-view geometry including image formation, epipolar geometry, projective transformation can be applied [53]. This may be in combination with the feature detection and matching strategies [54]. The concepts of space-time reasoning [55] and appearance-based matching remain more relevant for non-overlapping cases [56]. We work with the overlapping scenarios in this thesis.

In the context of multi-camera system, several camera network fusion strategies can be applied. The camera networks can be classified as centralized, distributed or clustered based on the fusion strategies [5]. Centralized processing configuration involves a single central fusion node which receives and processes all the camera

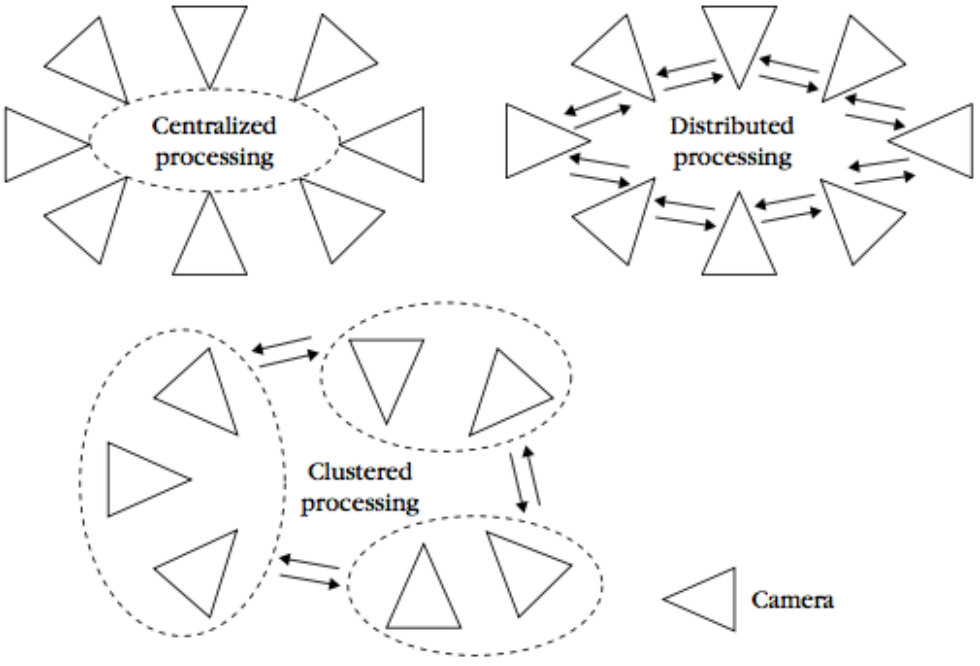


Figure 2.4: Fusion strategies in multi-camera networks. Reproduced from [5].

feeds. Whereas, in the clustered processing there are several local fusion nodes, instead of a single node, fusing the information together. There is no such fusion of processing occurring in the distributed processing technique. These concepts are illustrated in the Figure 2.4. For a survey of the techniques on distributed and decentralized processing the paper [57] can be considered. We utilise the centralized fusion processing strategy in this thesis.

Finally, we also introduce the concept of the wide baseline images. Baseline is termed as the line joining the projection centers between the two cameras. In case of multiple cameras, pair-wise combination between the cameras can be considered. We can also define the width of the baseline for cameras: it is the distance which cameras generate as this particular baseline is traversed. Cameras having wide baseline will generate varying looks across the views. Wide baseline images require less images than short baseline (such as stereoscopic rig) in terms of the FOV. Furthermore, wide baseline images provide better depth clarity and thus the FOV coverage. More details on wide baseline images can be studied in [25].

In terms of hardware: we focus our work on static, rectilinear camera networks. Nevertheless, there are also research applications using the dynamic and active vision sensors. Moreover, there is also a field of smart camera networks in which the architecture, middle-ware, and applications focus on remote, tiny, intelligent embedded systems. These topics and others on multi-camera vision can be visited in [5].

2.4 Multi-Camera People Detection

This section specifically focuses on multi-camera people detection, its applications for visual surveillance, and literature based on the improvement of detection. The challenges pertaining to people detection include the involvement of human articulations, scale and appearance based variations, occlusion, density, and environment clutter. Extensive research has been performed on the single camera human detection algorithms; however, these systems remain limited in their ability to handle occlusions, dense and cluttered environments [30]. As people detection is a well-studied issue of importance, finding applications in the domain of visual surveillance; thus, as a possible remedy, the research community has focused on using multi-camera systems for improvements in people detection and thus the

visual surveillance [13–15, 23, 24, 58]. The ubiquitous presence of cameras with the increase in computational resources has also fueled the development of multi-view research; sensor fusion, multi-view visual analysis are some of the challenges faced [5, 7]. Before we begin the presentation of the literature, a summary of monocular and multi-camera systems is presented in Table 2.1 and 2.2 respectively.

Fusion of information across multiple cameras requires a level of consistency across all the views of the object of interest. This fusion must also determine the presence or the absence of the object. Registration of an object present across multiple camera views can be used to estimate its location. One common approach in these systems constraints the search space to the ground plane using the planar world assumption [13, 15, 23]. Therefore, assuming the non-floating objects, planar homographies are calculated for the ground plane. The use of only ground plane may not be robust for several reasons such as bad foreground detections or the occlusion of the lower part of the body [12]. Recent approaches [13, 15] extend this by using multi-planar homographies combined with the ground plane.

Multi-view object detection can also be achieved with the aid of image registration with the use of camera calibration information of various cameras present in the scene. The use of camera calibration instead of camera homographies provides a more detailed scene reasoning to project the camera views to a common search space such as the ground plane [58, 59]. In addition to geometric techniques: probabilistic methodologies [23, 58, 59] have also been utilised for multi-view detection such as modeling people occupancy with primitives located on a discretized ground plane. Occupancy is the probability of the presence of a person at a particular position on the reference plane. We assume the reference plane as a real world plane to which all the camera views are projected or warped.

Khan and Shah [13] use the camera homography constraint to generate occupancy maps which is the fusion of projected scene planes across various views. The authors apply the multi-planar homographic constraint and combine it with graph cut segmentation in order to track people [13]. No calibration information is required but planar references must be present in at least one of the views and affine homography must be manually computed by the user for each sequence. Their proposed solution suffers from false positives or ghosts due to the limitations of the homography constraint. Khan and Shah account for ghosts using the space-time occupancies.

Eshel and Moses [60] perform people tracking in a dense, crowded environment using homography constraints applied only at the top layers combined with the pixel intensity correlation, motion direction, and velocity constraints. This method requires the use of partial calibration data. Temporal information is used to reduce phantoms. But, the algorithm is limited to those sequences in which heads are visible in a top view configuration. In summary, the algorithm uses multi-planar projections for top-view camera topologies in order to perform head detection.

Different from the two techniques mentioned above: Fleuret et al. [23] define a probabilistic occupancy map (POM) based on a quantized ground plane. They also define a distance measure in relation to the multi-view projections. Thus, POM is an algorithm to estimate an approximation of the marginal posterior probabilities of presence of individuals at different locations of an area of interest, given the result of a background subtraction procedure in different views. The camera calibrations are indirectly provided through a family of rectangles which approximate the silhouettes of individuals located at the considered locations. They further integrate it with Hidden Markov Model (HMM) for joint color, motion and occupancy modeling to perform tracking. The method suffers from high false positive rate and has a high computational cost. Moreover, this algorithm is limited to tracking up to a maximum of six people and fails to account for height variations. In [58, 59], the authors extend POM: they improve the localization accuracy by performing an optimization process that fits a cylinder instead of rectangle. The cylinder-fitting process approximating a person is achieved using the multi-planar projective features. However, this method is sensitive to the detection, localization of the feet.

The literature on multi-camera detection may also be classified into *direct* and *inverse* methods. Direct methods erect the objects from primitives. Examples of such primitives could be the segmented parts in the image or the silhouette blobs [61]. The inverse techniques assign a fitness value, a single figure of merit, to many possible object configurations. This is followed by an optimization process in order to find the best configuration corresponding to the objects. Examples of the inverse methods are the works presented in [58, 59, 62].

The majority of the methods presented so far suffer from a high false detection rate. For the geometric, multi-planar or homographic techniques, fusion of the projections corresponding to different people in the occupancy map could generate false detections, these detections of non-corresponding regions in the projected

space are referred to as the "ghosts". Besides focusing on the rather complete detection or tracking systems, this phenomenon of false detections has been studied in the literature as the ghost pruning problem.

Ren et al. [14] define ghosts as the false positives due to the intersections of non-corresponding regions. They propose to use color template matching for ghost pruning. But, as shown later, their method is unable to account for views with high variations in the color constancy. Moreover, their equations are limited to only two views. Unlike [14], we tend to propose works having no limitation in the number of views, number of planes used and that accounts for the views which lack color constancy.

Evans et al. [24] compute the probability of a ghost detection based on a spatiotemporal relationship of the objects present in the scene with the camera positions. The authors introduce a suppression map technique which is able to predict the possible location of the ghosts but it requires prior information about the location of the objects of interest which is obtained from the previous frames.

During this work, we try to achieve better detection performances without the use of any temporal information. The tracking of an unknown and variable number of people in a scene, without constraints, is still a challenge. Even if some methods do not use detection (e.g. [63]), other ones are based on the detection of people in fixed images(e.g. [13, 15]). The latter then perform tracking via association of these detections along time. As association is a complex task and as complexity increases with the number of objects detected. It is important to use, as input, as perfect detections as possible. To this end, the goal of this work is to detect people in still images, without considering temporal information in order to facilitate the tracking, or higher-level semantic, modules. Further, we have identified a core issue in multi-view geometry based detection, i.e. ghosts, and keeping this in mind, we try to achieve robust people detection.

In this thesis, we present a novel approach for pedestrian detection using multi-camera occupancy maps and by modeling the object shapes as 3D geometric primitives. We define a spatially varying kernel which depends upon the shape and geometric characteristics of the primitives and the camera calibration. We propose an analytical formation model for object detection by performing convolution of the proposed kernel with the object location map. Our spatially varying kernel is able to perform suppression of false detection through multi-view reasoning.

This specific kernel allows us to define sharp peak responses corresponding to the object detections. These detections can be localized by a deconvolution process. We also propose an efficient approximate deconvolution using a modified version of watershed transform specific to our kernel.

Further, we propose another novel method for people detection across multiple synchronized views based on coherence analysis. We have observed that it is possible to generate a shape model in the synergy map based on the location of an object in the scene. We refer to this shape as the Occupancy Shape Model (OSM). This shape model is a map created by modeling the person as the axis of a cylinder, at a given 3D location, followed by the fusion of the multi-planar projections of its synthetic images. We apply this model for ghost pruning using a similarity measure between the OSM and the real occupancy map. The 3D locations, at which our model is processed, are obtained by the application of local maxima detection using a modified watershed transform on the real occupancy map. Thus, our algorithm is based on the knowledge of the multi-view scene geometry. Finally, we perform cluster analysis on the similarity measures to automatically define the decision boundary for people detection.

The propositions account for challenging situations such as lighting, color, weather variations, and is able to robustly localize the pedestrians handling occlusion and projective shadows from the relatively dense scenarios. The propositions are not limited to a head-only camera topology [60], require no temporal information [24], account for color variations [14], perform multi-camera reasoning rather than the concatenation of monocular primitive detections [31], and have lower number of parameters with no optimization requirements as in [58, 59]. Finally, we propose the quantitative analysis, in comparison with the state-of-the-art [13, 23, 58, 59] to demonstrate the efficiency of our technique on popular public datasets [26–28].

2.5 Summary

We have presented a detailed survey of the related work and literature in this chapter. We began with the monocular people detectors to see their limitations and make an extension towards multi-camera visual surveillance. We presented the core vision-related technologies of the multi-camera systems. We then focused on calibration, topology and recognition in terms of multi-camera people detection.

Finally, we have presented the state-of-the-art in multi-camera people detection in link to our own work. A summary of the state-of-the-art has been summarized in Table 2.1 and 2.2. After a presentation of the goals of this thesis, we have briefly introduced the two propositions. These propositions are presented, in detail, in the following chapters.

Table 2.1: Summary of some monocular systems discussed in the related work.

Monocular Systems						
Method	**Category**	**Foundation**	**Strengths**	**Weaknesses**	**Relevance**	**Year**
Dollar et al. [30]	Literature Survey	Multiple techniques for people detection	-	-	Limitations of monocular detectors; need for multi-camera	2012
Yilmaz et al. [34]	Literature Survey	Multiple techniques for object tracking	-	-	Presentation of tracking as an additional module for surveillance	2006
Carr et al. [31]	Monocular people detector	Scene geometry, primitive modeling	Real-time, robust people detection using camera calibration	Multi-camera fusion, cases with decreased detection rates	Multi-view geometric method, potential for multi-view reasoning	2012

Table 2.2: Summary of some multi-camera systems discussed in the related work.

Multi-camera Systems						
Method	**Category**	**Foundation**	**Strengths**	**Weaknesses**	**Ghost Pruning**	**Year**
Aghajan & Cavallaro [5]	Book	Multi-camera works (Computer Vision & Sensor Fusion aspects)	-	-	-	2009
Wang [7]	Literature Survey	Multi-camera works (Computer Vision aspects)	-	-	-	2013
Khan and Shah [13]	Multi-camera people tracking	Homographic occupancy constraint (detection)	Occlusion	Ghosts (detection)	Temporal (tracking)	2009
Eshel and Moses [15]	Multi-camera people tracking	Homographic occupancy constraint (detection)	Density	Ghosts, topology	Temporal (tracking)	2010
Fleuret et al. [23]	Multi-camera people tracking	Probabilistic	Novel generative model	Speed, number of people, false detections	Not applicable	2008
Utasi & Benedick [58]	Multi-camera people detection	Probabilistic	Novel multi-view features, applied marked point process model	Sensitive to feet	Not applicable	2013
Ren et al. [14]	Ghost Pruning	Color features	Reduced false detections	Color variations	Color	2012
Evans et al.[24]	Ghost Pruning	Multi-view geometry	Reduced false detections	Suppression of known objects	Geometry, temporal (consecutive frames)	2012

Chapter 3

Ghosts in the Multi-camera Occupancy Maps

3.1 Introduction

Occupancy map is the fusion of projected information, present across multiple camera views, on to the common world coordinate system. Occupancy maps are calculated as the probability (or the confidence) of the presence of an object at the quantized ground location. These probabilities are determined by the projection of the camera views at the ground planes and several planes parallel to the ground planes, thus aggregating the evidence across several views and scene planes.

One popular application of occupancy maps is the detection of objects of interest employed in the multi-camera visual surveillance. Occupancy maps have been known to show robustness to occlusion, noise, variations of color or light. For this purpose, foreground masks are projected across the scene planes using camera calibration or by calculation of planar homographies. The method while having a see through effect, that is accounting for full or partial occlusion, suffers from a phenomenon of false detections which is referred to in the literature as ghosts. The objective of this chapter is to introduce the concept of occupancy maps and ghosts generation, so that we can proceed towards better handling, and achieving robustness to ghosts that is performing ghost pruning.

Figure 3.1 shows the block diagram for object detection using several overlapping cameras. In a first step, foreground masks are extracted from each camera view.

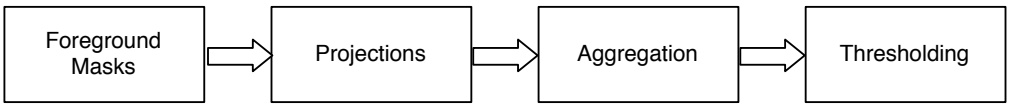


Figure 3.1: Block diagram for detection using occupancy maps

These images are then projected on planes parallel to the ground planes. All these projections are aggregated in a single map: the occupancy map that provides information about the presence of object. A simple threshold on this map is then sufficient to retrieve the objects of the scene. [1]

3.2 Multi-planar Projections and their combination

We present the multi-camera occupancy maps generated by the simultaneous aggregation of all the camera views across the various planes of the scene. The fundamental of this approach is that the projection of the image view to the ground plane, such as through homography, remains consistent across all the camera views. Moreover, this idea can be extended across several planar heights. In this section, we go through the process of multi-planar projections and also their synergy to generate the occupancy maps.

3.2.1 Foreground Masks

The first step to generate the occupancy map is the extraction of foreground masks in each camera view. It is generally done using background subtraction algorithms [13, 65, 66]. Instead of the Gaussian distributions [67, 68] as used by Khan and Shah [13] or Mixture of Gaussians (MoG) as employed by Utasi and Benedek [58, 59], our foreground masks are obtained using a multilayer background substraction method, based on Local Binary Pattern (LBP) and proposed by Yao and al. in [69].

Actually, even if methods based on Mixture of Gaussians (MoG) remain popular, they suffer from the balancing between the speed at which the model adapts to

[1] A part of this chapter appeared in the proceedings of the 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2014 [64].

the non-static background, and stability against the occluded background. The method proposed by Yao and Odobez [71] is a combination of photometric invariant color measurements, as a pixel-region model in the RGB space, and texture measurements using the local binary patterns, which provides structural information in the neighborhood of the pixel. This particular modeling of background allows to manage both texture and texture-less surfaces. This method is also able to handle non-static backgrounds thanks to a flexible weight adaptation in the background model. This is especially useful in moving areas such as the leaves of a tree or moving escalators present at airports, metro and train stations. The model is also able to account for addition or removal of stationary foreground objects e.g. a bag left in the train stations.

Local Binary Pattern (LBP) [72] is a texture-based descriptor. It is a gray-scale invariant statistic measure defined in a neighborhood around each pixel. The gray-level intensity (or the color) of each neighboring pixel of the studied pixel is thresholded, leading to a binary number (binary pattern) characterizing the texture around the central pixel.

We show the working of a basic LBP operator in figure 3.2. The texture operator is calculated on a 3x3 window around the central pixel. It assigns a label to each of the neighboring pixels by comparing them to the central pixel. The resulting binary assignment can be read in either clockwise or counter clockwise direction. LBP feature descriptor is a histogram designed with the decimal values of these binary assignments. Circular neighborhoods have also been proposed for LBP computation. One example is shown in figure 3.2 where P represents the sampling points defined on the circumference of a circle of radius R. The labeling criteria,

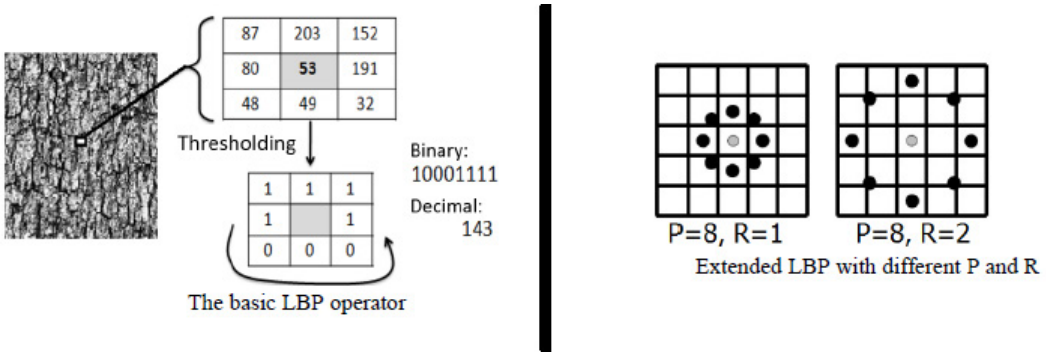


Figure 3.2: The basic and extended local binary patterns. Modified and reproduced from [70].

direction selection; including mathematical implementations for the different kinds of LBP feature descriptor are discussed in [72].

The main advantage of LBP is that it tolerates global and local illumination changes, and is computationally inexpensive. However, it fails for a texture-less object or when the textural description between background and foreground is the same. Yao and Odobez [71] thus add a RGB color descriptor to the LBP one. The background modeling thus consists of several modes m_k based on LBP and color features. Each mode is characterized by seven components $m_k = \{I_k, I_k^{min}, I_k^{max}, LBP_k, w_k, w_k^{max}, L_k\}$ where:

- I_k denotes the average RGB vector composed by the average of three colors R, G, B of the mode.
- I_k^{min} and I_k^{max} are the maximal and minimal RGB vectors that the pixels associated with this mode can take.
- LBP_k denotes the average of the local binary pattern learned from this mode
- $w_k \in [0, 1]$ denotes the weight factor, i.e. the probability that this mode belongs to the background.
- w_k^{max} represents the maximal value of w_k reached in the past.
- L_k is the background layer number to which the mode belongs. $L_k = 0$ means that the mode m_k is not a reliable background mode. $Lk = l > 0$ indicates that m_k is a reliable background mode in the l-th layer.

The detection then works as follows: for a new pixel characterised by its LBP and RGB values, the nearest mode m_k is searched. If it is close enough to the pixel (the distance is below a threshold), the mode is updated, otherwise, a new mode is created. The considered distance between a new pixel and a mode is a weighted average between color distance and texture distance. In order to be invariant to illumination changes such as shadows and highlights, the comparison between RGB values (color distance) is defined by their relative angle in the color space. Moreover, the weight w_k of all the modes are updated according to a novel 'hysteresis' scheme in order to manage moving background objects.

For the foreground detection, a distance map is built, similarly to the foreground probabilities map in the MoG method. Furthermore, the final results are smoothed using a bilateral filter applied to the distance map in order to decrease the noise.

The mathematical details of this background modeling, the learning process, the distance measures employed and the foreground detection are present in the original paper [69]. The authors have shown encouraging results in real and simulated datasets. The method is shown to produce fast and reliable results in surveillance scenarios on real databases obtained from a metro station, train station, and on general surveillance scenarios. We have utilized the default parameters for the generation of the binary foreground masks as given in the online available code [2]. A complete review of the background subtraction techniques in context of multi-sensor surveillance can be read in [73].

Figure 3.3 shows some examples of the foreground masks obtained. As it can be seen: some images are well segmented (see Figures 3.3(d) and 3.3(e)); in other cases, false detections can appear or some areas can be missed (see Figure 3.3(f)). It clearly appears that a multi-planes and multi-views consistency analysis yields results quality that depends on these different phenomena. The detection algorithm has to therefore manage imperfect foreground detection.

Next, the foreground map of all camera views are projected on several planes to create the occupancy map as explained in the following sections.

3.2.2 Multi-planar Projections

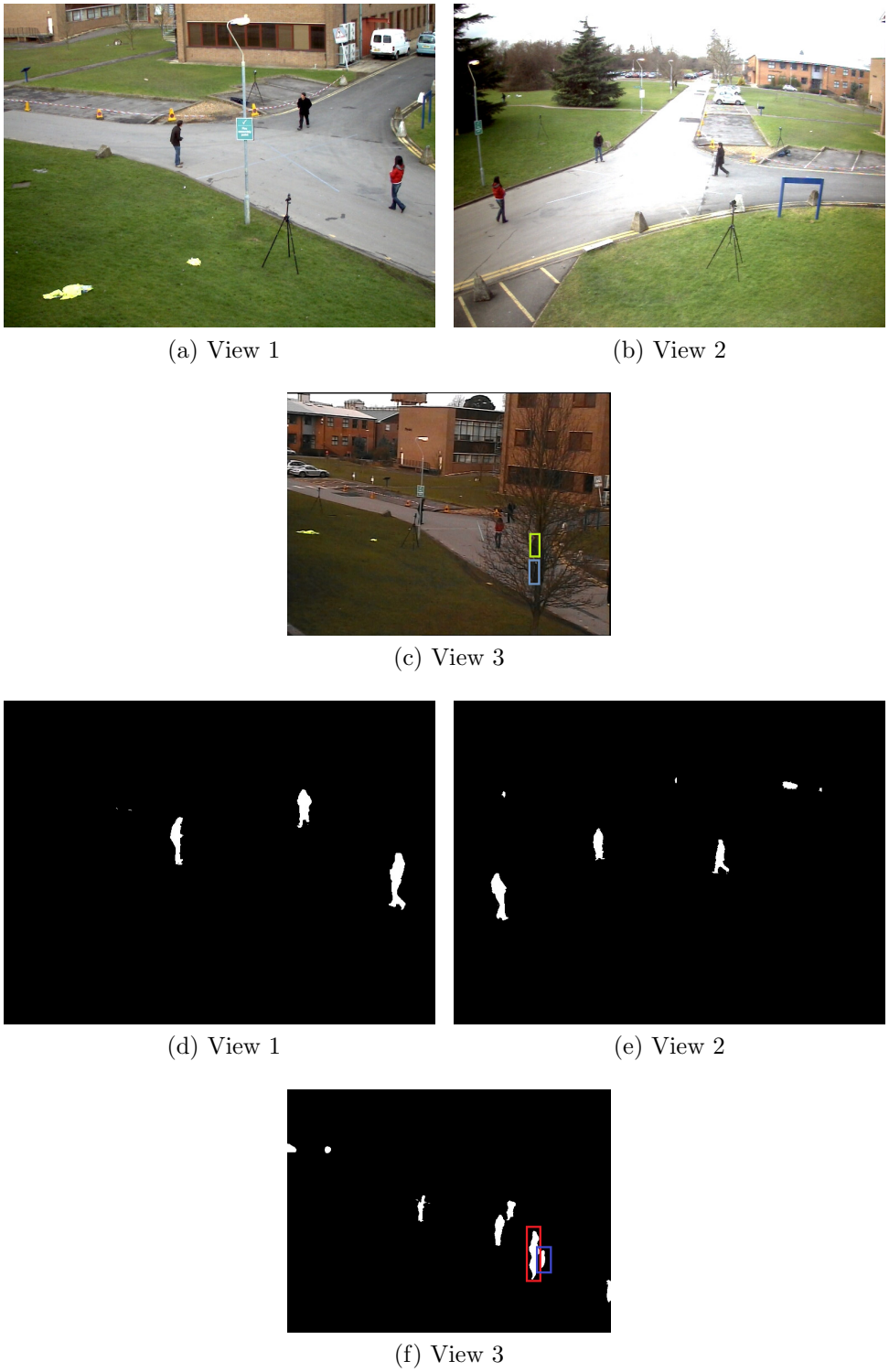
The camera calibration model is used to project the silhouettes obtained by background subtraction to the ground plane and several planes parallel to it.

Foreground silhouette maps are projected on the ground plane P_0 and on several planes P_z parallel to the ground plane at height z in the range of human height, as shown in Figure 3.4. These projections are made in two steps. First, the projection on the ground plane P_0 is obtained using the camera calibration [43]. Then, the projections on the parallel planes P_z are efficiently computed from the projection on P_0 using the following equations [58]:

$$x_z = x_0 + \frac{(x_c - x_0)z}{h_c} \tag{3.1}$$

$$y_z = y_0 + \frac{(y_c - y_0)z}{h_c} \tag{3.2}$$

[2] The foreground masks used in this thesis have been extracted using the default parameters of the code available at: http://www.idiap.ch/~odobez/human-detection/index.html



(a) View 1 (b) View 2

(c) View 3

(d) View 1 (e) View 2

(f) View 3

Figure 3.3: Example of a PETS 2009 frame and its corresponding foreground masks. The bounding boxes in (c) are the approximate locations of two persons. The red bounding box in (f) points out one missed detection and the blue bounding box is for the false detection.

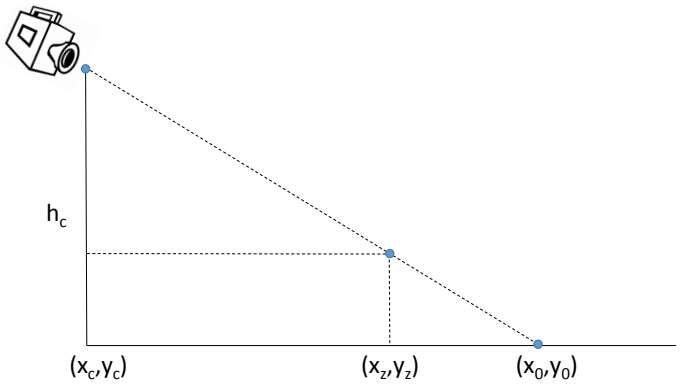


Figure 3.4: Projection of point of the image on a plane at height z and the ground plane.

where:

- (x_c, y_c) represents the camera position located at height h_c.
- (x_z, y_z) is the projection of a point on the plane P_z knowing that this point has been projected at (x_0, y_0) on the ground plane P_0

If the camera calibration model or homography information is not available then the method proposed by Khan and Shah [13] in section 4.2 of their paper can be used.

The same projection is applied for all the pixels of the silhouettes of the foreground map. Moreover, several projections are obtained for each foreground, one for each plane height z. As it can be observed on Figure 3.5 from top view, these projections are not located on the same position. The projected shape is scaled and translated with regards to the height z of the plane projection.

Figure 3.6 illustrates the same process obtained on a randomly selected frame of the real sequence PETS 2009. Three projections are presented: on ground plane and two parallel planes at height 100cm and 190 cm. It can be noticed that, as the size of the area of projection is fixed and is the same for the different heights, some object can be visible for a height and not for others. It is the reason why some objects not present in Figure 3.6(a) are visible in Figures 3.6(b) and 3.6(c).

These projections, detailed for one camera, can be done and merged in a multi-camera context to manage occlusion.

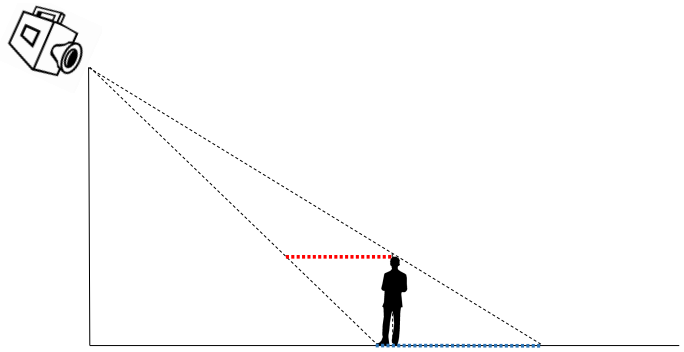
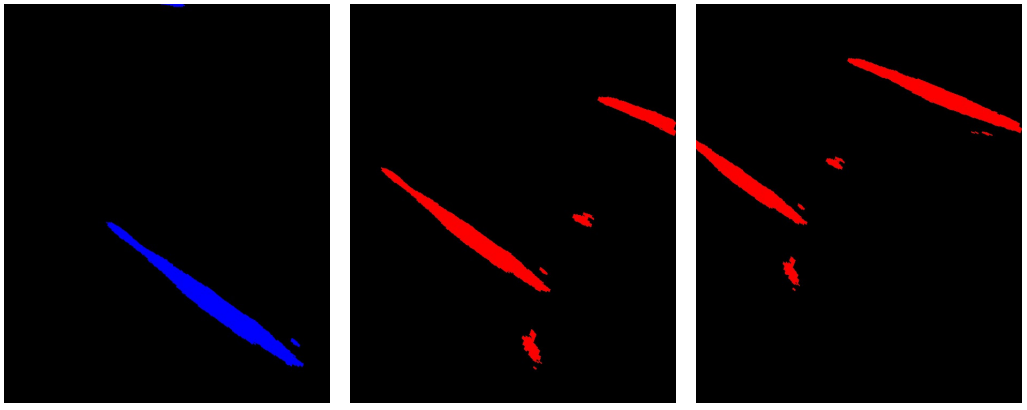


Figure 3.5: Projection the people perceived by the camera on the ground plane (blue) and a parallel plane at height z (red).

3.2.3 Homographic Occupancy Constraint

In this section we introduce the homographic occupancy constraint as defined by Khan and Shah [13]. Let us consider a scene with a single object observed by several wide-baseline stationary cameras as shown Figure 3.7. This object produces a single foreground region in each view. Let us define some notations:

- π is a reference plane, for example the ground plane or a plane parallel to it, in the 3D space ;
- Φ_i is the foreground region in view i;
- $H_{i_\pi j}$ is the homography induced by plane π from view j to view i;



(a) (b) (c)

Figure 3.6: Multi-planar projections on a frame of PETS 2009. Projections at (a) ground (b) 100 cm (c) 190 cm.

- P is a point of the space that is projected in p_1, p_2,..., p_i in any n view.

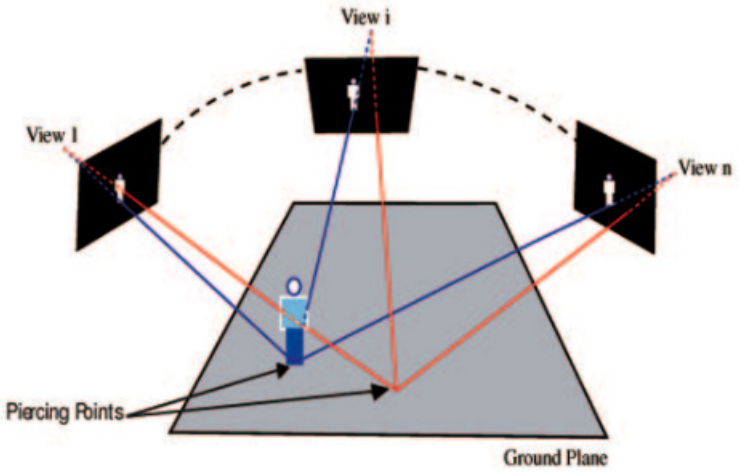


Figure 3.7: Principle of multi-views projection. Reproduced from [13].

With these notations, Khan and Shah [13] introduce two propositions:

Proposition 1: If P lies in the reference plane π and is inside the volume of the object, then

- p_i, its projection in view i belongs to Φ_i
- $p_i = [H_{i_\pi j}]p_j$

This proposition is illustrated on the same example than in [13] in Figure 3.7. Considering the feet of the people and the ground plane as reference plane, these feet are projected in a foreground region in all views. On the contrary, a point of the reference plane π, outside of the volume of the object will be projected on foreground or background area, according to the views.

The homographic occupancy constraint is defined as the following proposition:

Proposition 2: Consider a particular view j with its foreground region Φ_j. Then, $\forall p_j \in \Phi_j$, $p_i = [H_{i_\pi j}]p_j$ belongs to Φ_i.

This second proposition assures the management of occlusion as explained in Figure 3.8 based on the same example as that in [13]. Actually, in view 1, the green person is occluded by the blue one and its feet are not visible. It results in the fusion of both foreground regions in a single region. However, all these pixels satisfy

the homography occupancy constraint. On the view 2, both persons are visible. If $H_{1_\pi 2}$ is the homography induced by π from view 2 to view 1 then for all p_2 of the view 2, $p_1 = [H_{1_\pi 2}]p_2$ belongs to the single silhouette in the view 1.

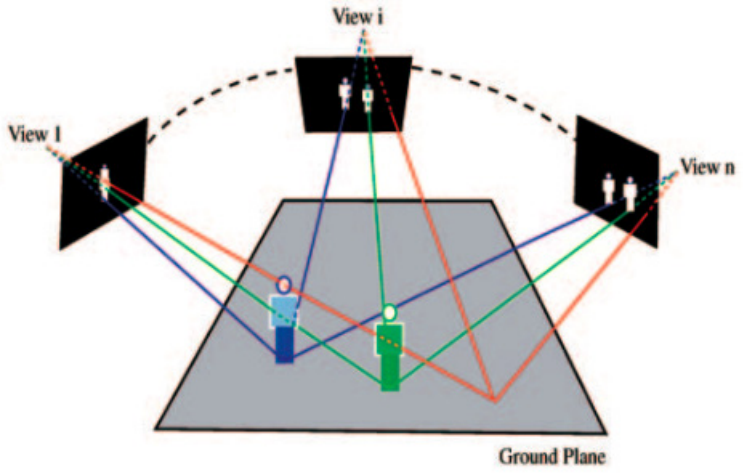


Figure 3.8: Principle of multi-views projection in case of occlusion. Reproduced from [13].

The occupancy map gathers the information to measure, for all position P, how much the homography occupancy constraint is satisfied. It is computed by merging all the projections calculated for all the views to detect people and to deal with occlusion problem.

3.2.4 Multi-camera Occupancy Maps and Ghosts effect

Khan and Shah [13] define the occupancy map as a 2D grid of object occupancy likelihoods. Occupancy map is obtained by the fusion of the multi-planar projections[3] previously defined and calculated for several cameras. Let us assume that $\mathbf{P_{v,P_z}}(x, y)$ is the map corresponding to the projection of the camera view $v \in V$ on the plane $P_z \in P$. V is the set of views and $n_v = card(V)$ is the total number of views i.e. the number of cameras of the system. Then, the occupancy map $\mathbf{O_{P,V}}(x, y)$ is defined as follows:

$$\mathbf{O_{P,V}}(x, y) = \frac{\mathbf{S_{P,V}}(x, y)}{max(\mathbf{S_{P,V}}(x, y))} \tag{3.3}$$

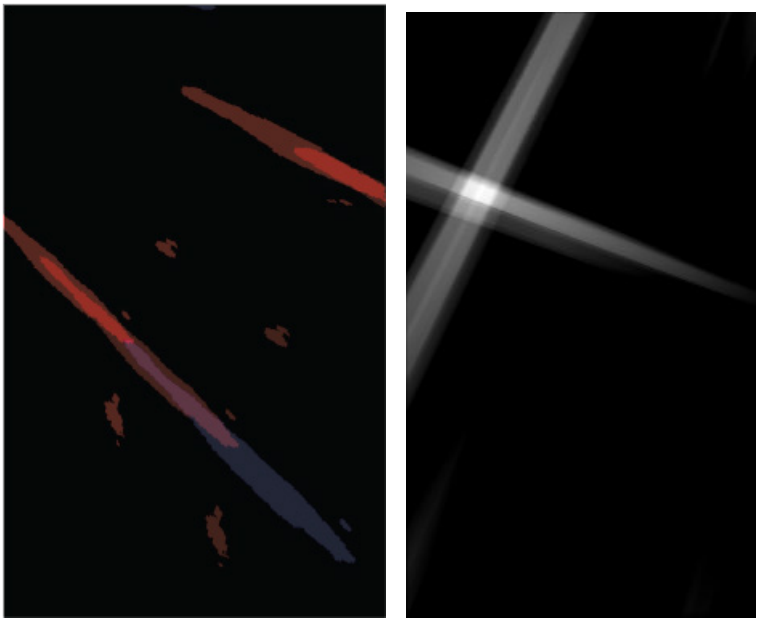
[3]The code for multi-planar projections is available at: http://web.eee.sztaki.hu/~ucu/mbd/sw/fgprojection.tar.gz.

where $\mathbf{S_{P,V}}(x, y)$ is the accumulation map defined by:

$$\mathbf{S_{P,V}}(x, y) = \frac{1}{n_v} \sum_{P_z \in P} \sum_{v \in V} \mathbf{P_{v,P_z}}(x, y) \tag{3.4}$$

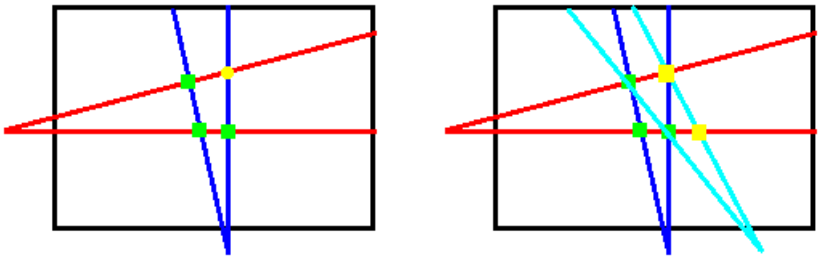
Figure 3.9(a) and 3.9(b) show the combination of the projections of respectively a single and two camera views. This combination is done for the three heights presented figure 3.6.

At each location of the occupancy map is assigned a value depending upon the number of cameras and the number of planes. Ideally, a nonzero value appear only if the point is an element of the foreground mask. Let us consider the case of one plane projection, and more specifically the ground plane, then the occupancy map value will be the highest if the feet of a person are visible across all cameras. The value should decrease with the number of cameras observing the corresponding point. If the foreground is partially segmented, only the plane projections of the detected parts will contribute to the values of the occupancy map. The use of several planes (from the ground up to the head) tackles the imperfection of the foreground. Thus, the values of the occupancy map are directly linked to the number of planes and views on which a foreground is projected. The highest value



(a) (b)

Figure 3.9: Illustration of the combination of planar projections used in the Figure 3.6. (a) is for one single camera. (b) is for two camera views.



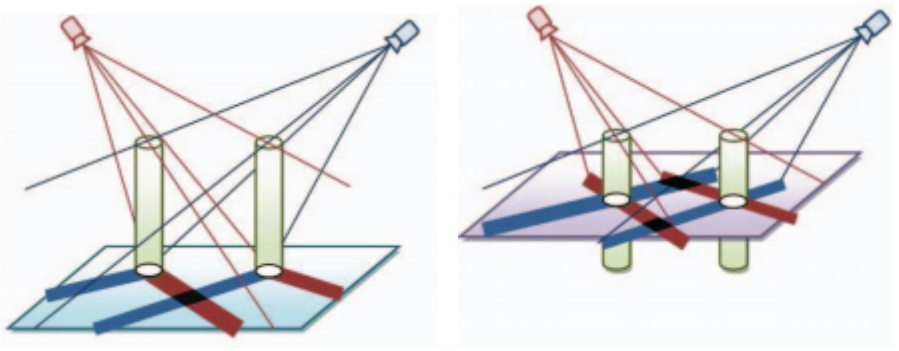
(a) (b)

Figure 3.10: Illustration of ghosting phenomenon in a schematic scenario. Black boundary represents the physical area. Red, blue and cyan are lines drawn from 3 cameras centers to the known objects. Green points are the real objects and yellow points are ghosts. (a) illustrates ghost generation using two cameras. (b) is generated using an additional camera with a limited Field of View (FOV).

is obtained for a foreground visible across all cameras and across all the parallel planes.

In this ideal case, an object could be detected by locating the peaks in the occupancy maps as proposed in [13]. This method, however, suffers from false positives known as ghosts and due to the multi-views projections of several nearby objects. Ghosts produced by a chance alignment of projections coming from several objects are illustrated in Figure 3.10. It shows how the intersection of a line from the camera centre to the object of interest generates a ghost [24].

Let us consider another schematic example proposed by [14] and illustrated in Figure 3.11. Suppose that there is a person visible in the scene and that this person can be approximated with a tall cylinder. This cylinder is visible as a circle from the top view. From the side-views the cylinder is visible as a quadrilateral. The projection of this cylinder on the ground plane produces a large line which extends from the lower end of the cylindrical centre to the optical centre of one camera. Besides the position of the two centers, the line also depends upon the plane on which the projection occurs. The projected line moves away from the camera as we approach lower level planes such as the ground. Similarly, the line moves away as we increase the planar height. This gives rise to the star shaped structures in the scene. When the cameras are imperfectly aligned, the number of streaks of one star is equal to the number of cameras.



(a) Ground Plane (b) Parallel Plane

Figure 3.11: Illustration of ghosting phenomenon across different planes. The ghost problem occurs across several planar heights of the scene. Reproduced from [14].

Figure 3.12 shows another example of an occupancy map computed on real data. As it can be observed, a simple thresholding produces various false detections represented by the green localizations. Around each real position of people (blue point), we observe star shaped structures with three streaks corresponding to the three cameras.

Authors have proposed to tackle the ghost problem by using temporal reasoning and space [23, 60]. Alternative solutions to these complete detection/tracking system have been proposed. For example, Ren et al. [14] define ghosts as the false positives due to the intersections of non-corresponding regions (see Figure 3.11). They propose to use color template matching for ghost pruning. Evans et al. [24] introduce a ghost removal technique which is able to predict the possible location of the ghosts based on the scene geometry. The suppression map assigns a probability of ghost detection based on the distance of objects from camera centers and camera calibration parameters. The method requires prior information about the location of the objects of interest that is obtained from the previous frames.

In the following chapters, we will study the possibility of improving per-frame detection accuracies and to remove the ghosts. To this end, we study the use of 3D geometric primitives and their respective shapes in the occupancy map space. This combination of 3D multi-view geometry and pattern recognition techniques helps for people detection.

3.3 Summary

In this chapter, we have introduced the concepts of multi-camera occupancy map, detection in the multi-camera occupancy maps, and the problem of false detections or ghosts which occur in such maps. Occupancy maps provide a reliable way to perform multi-camera people detection. The occupancy maps used in the thesis are generated using the aggregation of multi-planar scene projections across several heights of the scene. Occupancy maps provide robustness to complex situations such as occlusion handling, lighting variations, shadows [13, 60].

In the context of object or people detection such as for multi-camera surveillance, this popular technique of multi-camera occupancy maps suffers from various false detections. This phenomenon is referred to as ghosts in the literature [14, 24]. Ghosts occur for particular configurations of the objects and cameras in the scene where intersections appear.

This thesis proceeds by presenting two methods based on the multi-camera occupancy maps for ghost removal. The first method focuses on the study of the particular shape generated around each object in the occupancy map space, and use it in order to perform people detection and ghost suppression. The second

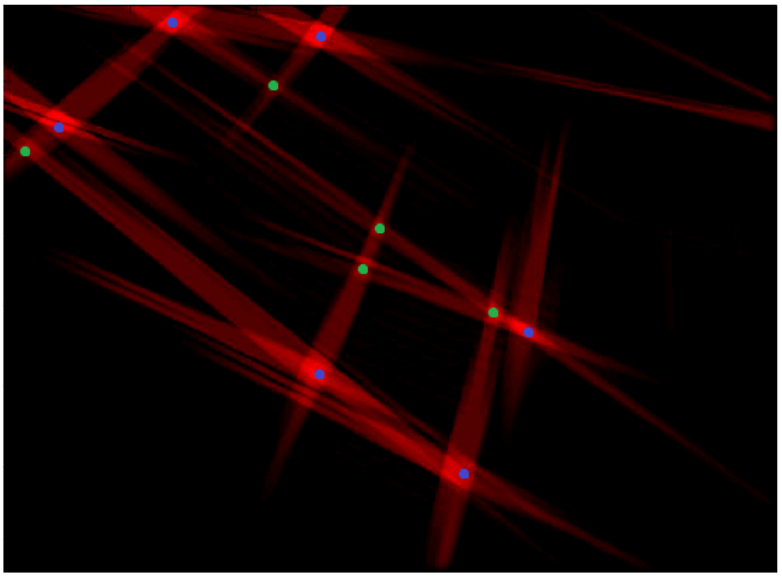


Figure 3.12: Illustration of an occupancy estimated on real data. Brighter red colors indicate higher probabilities. The blue dots represent the ground truth that is the location of the people. The green dots represent the ghosts.

method identifies key locations in the occupancy map and then performs reasoning around ghost pruning in order to achieve people detection. It is relevant to point out that as we are focusing on detection therefore we further constraint ourselves not to use prior knowledge about the number of people or their locations. As shown later in the thesis, these two techniques are able to account for robustly detecting the people, reducing the number of false detections, and are not limited in the number of views or number of planes used. They are also able to account for views such as with the lack of color constancy, shadows, occlusions - without using temporal information.

Chapter 4

Multi-camera Occupancy Map Deconvolution

4.1 Introduction

In this chapter, we present an approach for multi-camera people detection exploiting the concepts of multi-views geometry and the shapes of 3D geometric primitives. Multi-camera occupancy maps provide peak responses corresponding to the object detection but suffer from several false detections known as ghosts. The novelty of the technique in this chapter is the introduction of shape patterns which can model the objects, such as people, by defining a kernel function in the projected occupancy space. This kernel depends upon the geometry of the 3D primitives and also varies in relation to their position with respect to the cameras in the real world configuration. For multiple objects visible across several cameras, we define a formation model of the occupancy map which is the convolution of this spatially varying kernel with the set of possible object locations. The locations corresponding to detections can thus be obtained through a deconvolution process. For computational efficiency, we further propose an estimated deconvolution process specific to our kernel responses which can also be heavily parallelized. We evaluate the application of this process towards people detection by studying various 3D cylindrical primitives. Experiments on three public dataset sequences, presented later in this thesis, including a comparison with other approaches, show the efficiency of the proposed method in terms of people detection and ghost pruning, including in adverse and challenging conditions. Figure 4.1 illustrates

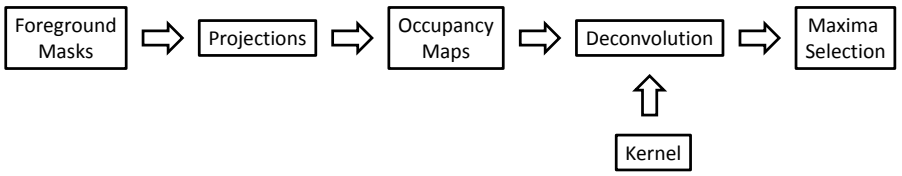


Figure 4.1: Block diagram for primitive detection process applied on the occupancy maps

the various steps through which the proposed algorithm proceeds. We discuss these steps in detail, later in this chapter; but; first, we present the modeling of the occupancy maps.[1]

4.2 Modeling of the occupancy map

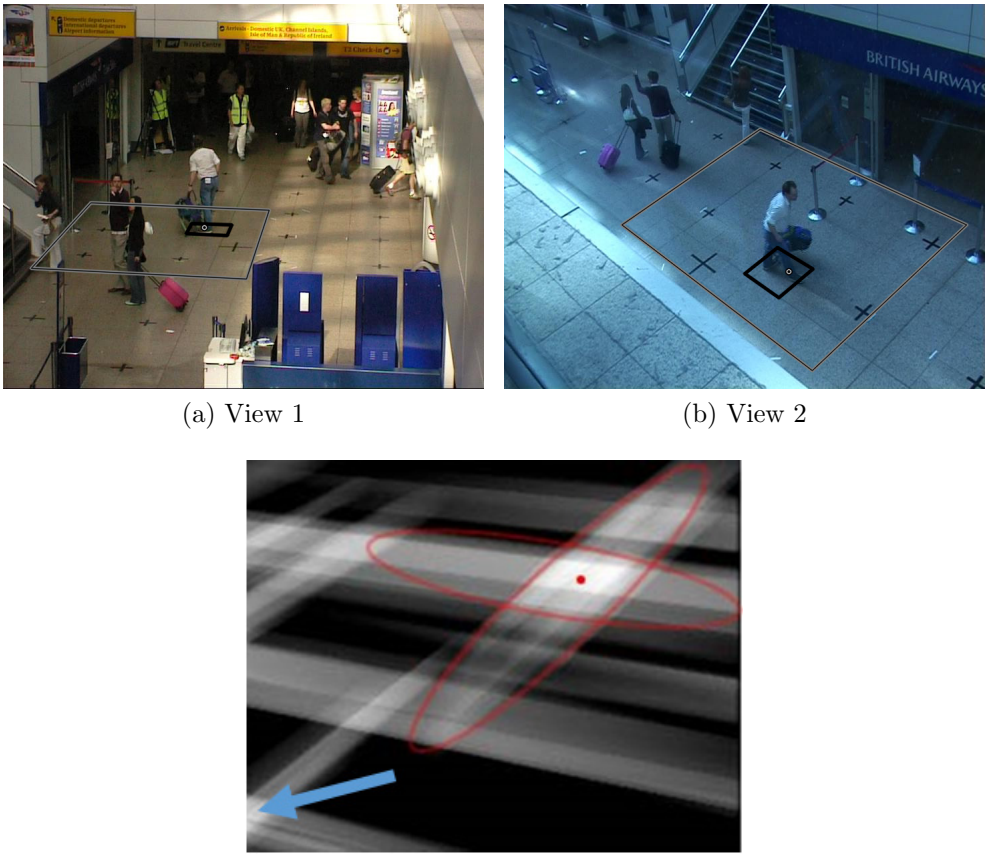
Occupancy maps are well known in the multi-camera context to exhibit peak responses corresponding to the object locations in the scene as shown in the previous chapter and illustrated in Figure 3.12. Multi-camera occupancy maps assign a probability that is based on the normalized sum of the image evidence, binary or probabilistic, gathered from all the cameras and projected to a common search space such as the ground plane.

Figure 4.2 shows an example of occupancy map generated using two camera views and multi-planar projections parallel to the ground. It can be observed in the figure that there is a star shape centered on the objects. The legs of this star correspond to the number of cameras observing the scene (two in this case). The rectangular bounding box in Figure 4.2(a) and 4.2(b) represent the Area of Interest (AOI) on which the occupancy map is computed (Figure 4.2(c)). It can be noticed in this figure that difficulties appear on real sequences, such as those arising from errors in the background subtraction process; and, the projection/presence of people outside AOI, but, present in the occupancy.

[1]This chapter appears as a section of a publication accepted in the proceedings of the 12th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), 2015.

The method proposed in this chapter is based on the modeling of the occupancy map generation that explains the particular star shape present around this object. This allows to derive the proposed detection method.

Let us consider the section 3.2.4 about the multi-camera occupancy maps and estimation. Let us suppose that an arbitrary 3D shape is present on the ground plane. It produces foreground masks in each camera view i which are projected on the occupancy map, as explained in equation 3.3. We can model the presence of a 3D object in the occupancy map. This object can be a cylinder, cube, pyramid or another shape. The selection of shape depends upon the object to be detected. At each location in the world, a 3D object on the ground plane X_o, Y_o corresponds to a specific occupancy map, referred as the kernel. To this end, the



(a) View 1

(b) View 2

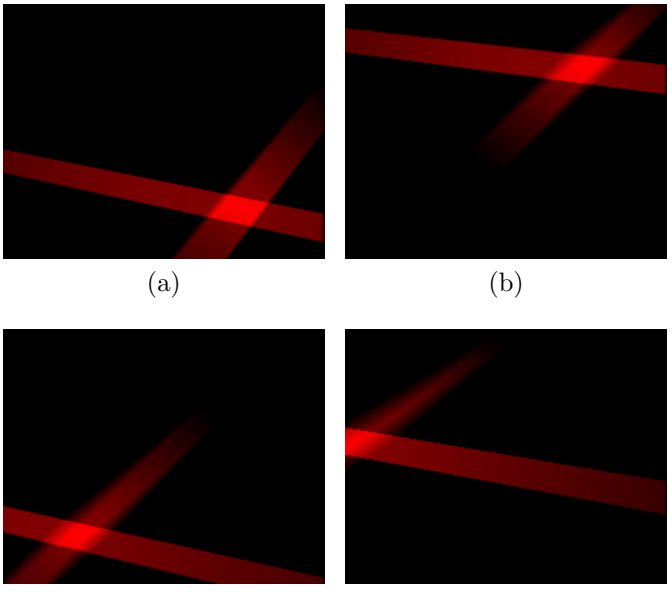
(c) Multi-camera Occupancy Map

Figure 4.2: Illustration of the multi-camera occupancy map and primitive based detection using two views of the PETS 2007 dataset. The occupancy map contains a star pattern corresponding to the person. The arrow points out a case of projection from outside the Area of Interest (AOI) affecting the occupancy map.

presence of an object at position (X_o, Y_o) on the ground plane generates a kernel $\mathbf{K}_{\mathbf{X_o},\mathbf{Y_o},\mathbf{P},\mathbf{V},\mathbf{S},\theta}(x, y)$, defined in the same Area of Interest (AOI) as that of the occupancy map, depending on:

- X_o, Y_o, the position of the 3D object,
- θ, the orientation of the object,
- P, the set of planes P_z used to compute the occupancy map (equation 3.3),
- V the different camera views, including camera configuration, geometry,...
- S the shape of the object

For this particular work focusing on people, primitives are selected as cylinders. As this shape is symmetric according to the vertical axis, and as we only consider cylinder in this application of people detection, the notation of the kernel can be simplified with $\mathbf{K}_{\mathbf{X_o},\mathbf{Y_o},\mathbf{P},\mathbf{V}}(x, y)$. Figure 4.3 shows four kernels generated at different locations (X_o, Y_o) of the 3D cylinder and two cameras. These kernels are computed offline.



(a) (b) (c) (d)

Figure 4.3: Example of kernel generated at different locations for a 3D cylinder and 2 cameras

Now, suppose that several people are standing on the ground floor. We can model their presence by a 2D function $\mathbf{D}(x, y)$ which is the summation of 2D Dirac delta functions $\delta(x, y)$:

$$\mathbf{D}(x,y) = \sum_{i=1}^{n} \delta(x - X_i, y - Y_i). \tag{4.1}$$

where (X_i, Y_i) are the coordinates of the n people. Ideally, the occupancy map, that reflects the probability of presence for each location, should be equal to this 2D function $\mathbf{D}(x,y)$. But as shown earlier, the process involved in its creation induces particular shapes around the people; these shapes called kernel as introduced before. So, the occupancy map can be mathematically modelled as the summation of convolution between the Dirac functions and the kernels:

$$\mathbf{O_{P,V}}(x,y) = \sum_{i=1}^{n} \delta(x - X_i, y - Y_i) * \mathbf{K_{X_i,Y_i,P,V}}(x,y) \tag{4.2}$$

4.3 People Detection

4.3.1 Deconvolution

Using this formulation, people detection consists in finding the different locations $\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$, knowing the occupancy map $\mathbf{O_{P,V}}(x,y)$ and the kernels $\mathbf{K_{X_o,Y_o,P,V}}(x,y)$ for all positions (X_o,Y_o). This can be achieved by the deconvolution of $\mathbf{O_{P,V}}(x,y)$ followed by a peak extraction process. However, as the shape of the kernel changes according to the position, we have to make a deconvolution of the whole image for each pixel, just changing the kernel used for the deconvolution, and keeping only the result for the considered pixel. This step is too computationally expensive to be implemented and thus approximate solutions have to be found. Even if correlation seems to be appropriate to solve this problem, we introduce another measure based on the intersection of the shapes and inspired from the intersection of histograms [74]. Assuming that the scene is not of overly dense crowds, this measure is utilized as an estimation of the deconvolution:

$$\hat{\mathbf{D}}(X,Y) = \frac{1}{\|\mathbf{K_{X,Y,P,V}}(x,y)\|_{max}} \sum_x \sum_y min(\mathbf{K_{X,Y,P,V}}(x,y), \mathbf{O_{P,V}}(x,y)). \tag{4.3}$$

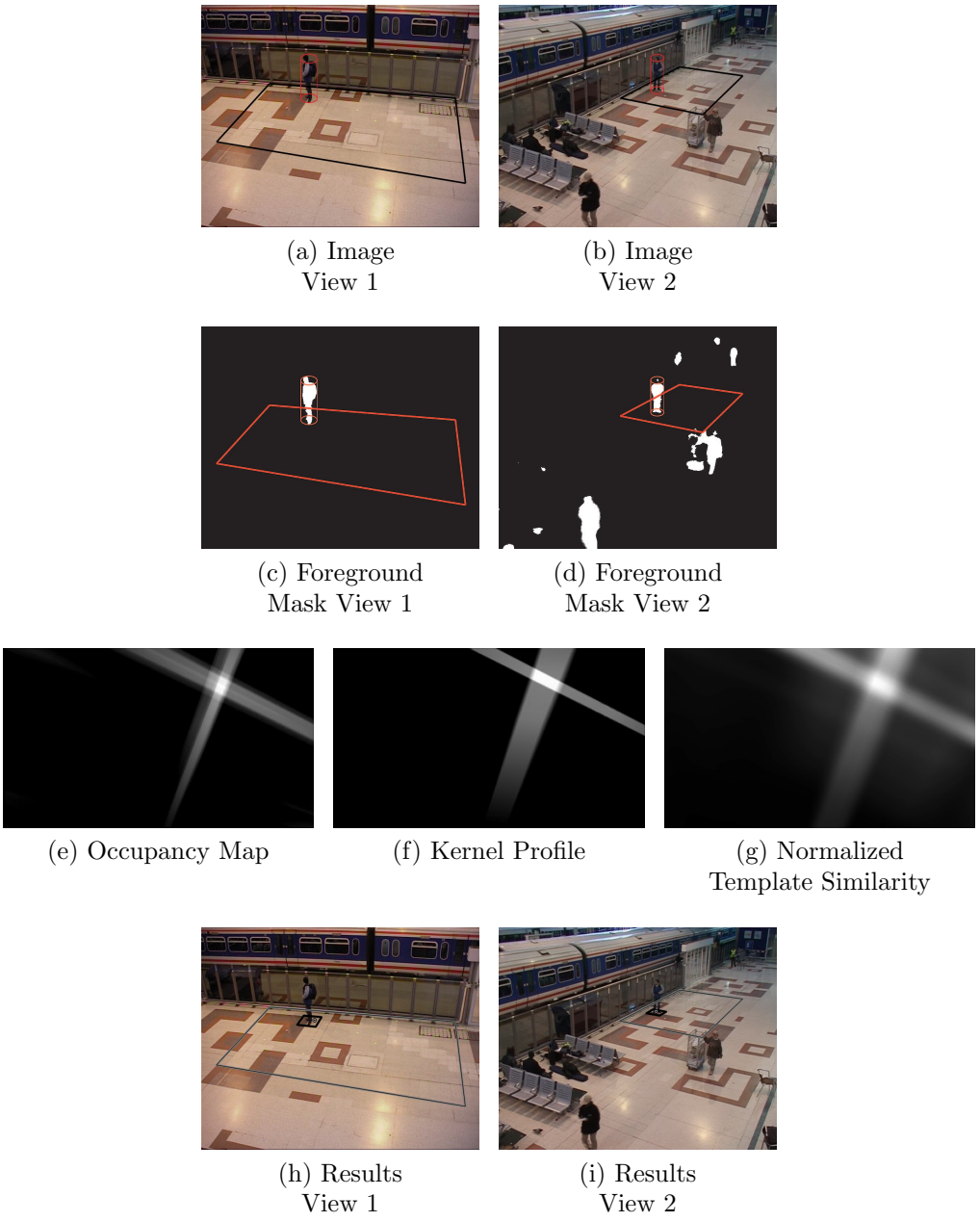
where $\|\mathbf{K}_{\mathbf{X},\mathbf{Y},\mathbf{P},\mathbf{V}}(x,y)\|_{max}$ is the max-norm [75]. $\hat{\mathbf{D}}_{\mathbf{K}}(X,Y)$ searches for any evidence of local matching and proceeds further by normalizing it with respect to the global kernel space. Actually, the size of the kernel $\mathbf{K}_{\mathbf{X},\mathbf{Y},\mathbf{P},\mathbf{V}}(x,y)$ changes according to the studied location (X,Y). This measure is further justified in section 5.3 and its utilisation is validated in Table 5.1. There exists a trade-off between the detection accuracy and the computational processing defined by the number of samples over which the similarity is computed. Higher number of samples produce a sharper response at the cost of the time required. This measure is quantitatively analyzed in section 5.5.1. Its principal advantage is to manage the case where several people are present in the scene.

By detecting in this way, we manage directly all problems by detecting people and avoiding ghosts:

- at a location where a person is present, the shape in the occupancy map and the shape of the kernel are similar. Thus an important value of $\hat{\mathbf{D}}(X,Y)$ is obtained.
- at a location where no one is present, there is no shape in the occupancy map. Thus a small value of $\hat{\mathbf{D}}(X,Y)$ is obtained.
- at a location where a ghost is present in the occupancy map, we suppose that the shape in the occupancy map and the shape of the kernel are not similar leading to a small value of $\hat{\mathbf{D}}(X,Y)$.

Figure 4.4 shows the whole process on an image of the PETS 2006 sequence with first the original images of both cameras (a) and (b). Then the foreground masks for each view are presented as well as the Area of Interest projected on this view (c) and (d). Figure (e) shows the real occupancy map computed from foreground images. Figure (f) presents the kernel computed with a 3D cylinder localized at the position of the people (the position has been manually selected for illustration. In the algorithm, a kernel is computed for all location). Figure (g) presents the result of the approximate deconvolution estimated using a similarity measure as explained equation 4.3.

A final step is necessary on the deconvoluted image: the selection of local maxima that leads to the positions of the people $\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\}$ in the scene.



(a) Image View 1 (b) Image View 2

(c) Foreground Mask View 1 (d) Foreground Mask View 2

(e) Occupancy Map (f) Kernel Profile (g) Normalized Template Similarity

(h) Results View 1 (i) Results View 2

Figure 4.4: Illustration of the various stages of the proposed algorithm. (Top two rows) Two views of PETS 2006 dataset, and their corresponding foreground masks. The boundaries represent the Area of Interest (AOI). The person is modeled by a 3D cylinder. (Third row) The occupancy maps obtained from the foreground masks, the kernel profile of the cylinder generated using the camera calibration, and the corresponding estimated deconvolution. (Bottom) The results obtained. The bounding box around the person represents the ground truth, and the circle marks the estimated detection. The proposed analytical model induces a maximum response for the object centre and the estimated detection is the result of this maxima selection in the estimated deconvolution.

4.3.2 Maxima Selection using Watershed

The template similarity image $\hat{\mathbf{D}}(X, Y)$ does not resemble to the required combination of Dirac delta responses (see Figure 4.4(g)). The result is a distribution consisting of several modes for which it is necessary to estimate the maxima to obtain the number of objects or people, and their locations on the ground plane. This step of maxima extraction is done using the watershed principle as it presents some advantages in this application.

Watershed is a concept in the field of geography - it defines the ridge which separates the regions drained by two separate rivers. This ridge is called the watershed line. The two separate drainage entities formed are called the catchment basins. The watershed transform applies this reasoning to the domain of image processing. It is popular for the problems of image segmentation.

Algorithm 1 Watershed based maxima selection

- Calculation of the local maxima in 8×8 pixel neighborhood blocks
- Sorting of the local maxima in descending order
- Perform flood-fill algorithm for each local maximum corresponding to the tolerance threshold
- Maxima in the already filled regions are discarded
- If several maxima occur in a flood-fill region; then, calculate the geometric center

In order to understand the watershed transform, we need to imagine the two-dimensional grayscale image in the form of a topological surface. In this topological surface, the height is defined by the grayscale values of the image. Let us imagine this topological surface as in the Figure 4.5. Further, we hypothesise rainfall over the 3D surface. The water falling on this surface is collected in the two catchment basins. The watershed transform locates the watershed lines and the catchment basins in the grayscale image. This then solves problems such as the mode retrieval or the segmentation of the images. But it is necessary to define a way in which the grayscale image can be represented in the form of a topological surface.

If we look at the Figure 4.4(g), and imagine it as a 3D surface (see Figures 4.5 and 4.6), it looks like a mountain. But by inverting this surface, a catchment basin

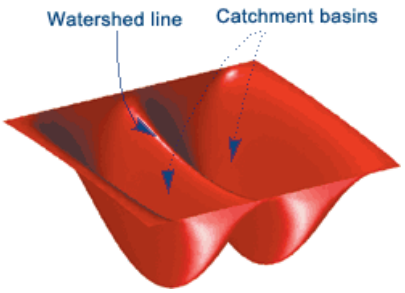


Figure 4.5: The catchment basins created with the watershed transform process. Reproduced from [76].

appears at the location of people. So, watershed theory can be applied to solve our problem of maxima detection and, more particularly, a modified version of the watershed transform with markers [77].

Keypoints are first initialized at the local maxima of the template similarity image $\hat{\mathbf{D}}(X, Y)$ calculated using 8×8 pixel blocks. These local maxima are sorted in descending order and inverted to define the markers (Figure 4.6). Flood-filling is then performed with these markers that are kept only if their topographical

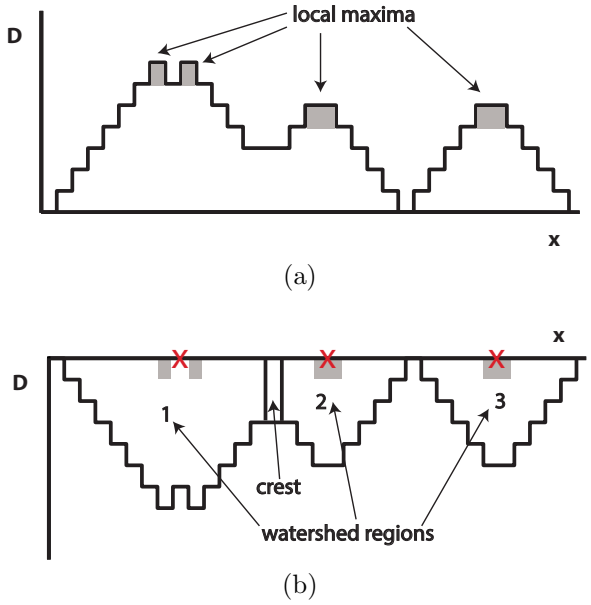


Figure 4.6: One-dimensional illustration of the keypoint extraction based on the watershed transform applied to the local maxima. (a) Local maxima are extracted on an arbitrary distance field D. (b) Inverted local maxima are treated as markers on which the marker based watershed transform is applied. The crest here acts as the watershed line and is defined by the τ parameter. The local maxima must be the greatest value in its region. In case of multiple similar local maxima, the geometric centre is taken as the keypoint location.

Algorithm 2 People detection by deconvolution

for all X of the occupancy map **do**
  for all Y of the occupancy map **do**
    compute the local kernel $K_{X,Y,P,V}(x, y)$ (parameter: $\varnothing$)
  end for
end for
for all time k **do**
  compute the occupancy map $O_{P,V}(X, Y)$
  for all X of the occupancy map **do**
    for all Y of the occupancy map **do**
      compute the template similarity i.e. $D(X, Y)$ using Equation 4.3
    end for
  end for
  local maxima using watershed, removing multiple ambiguities (parameter: τ)
end for

prominence is greater than a tolerance threshold τ. τ accounts for the maximum distance between two maxima to conclude that they are from the same watershed regions. So, the number of detected keypoints decreases as τ increases. Further, if multiple similar local maxima exist in one catchment basin then we define the resulting keypoint at their geometric centre. The τ value is directly linked to the number of watershed regions that are detected. By increasing the value of τ the watershed regions are likely to be fused. A 1D illustration of this process is shown in Figure 4.6 and the whole procedure is summarized in Algorithm 1. The implementation of the algorithm is publicly available[2].

4.4 Experimental Results

The whole method introduced in this chapter is referenced as MOD in the following, for Multi-camera Occupancy map Deconvolution. Before presenting the results, let us introduce the studied database and the evaluation measures.

[2]The executable is available at: http://imagej.nih.gov/ij/download.html. The find maxima plugin with 'point selection' output type is used.

4.4.1 Dataset

We have used a subset of *City center* sequence from PETS 2009 dataset [28] as defined in [58, 59]. The evaluation sequence contains 400 outdoor scene images obtained from three camera views (see Figure 4.7) and representing approximately 1 minute of video. We use the same scenario as [58] and, from the available views we selected cameras with large fields of view (Cameras 1, 2, and 3). We use the Area of Interest (AOI) of size $12.2 \times 14.9\ m^2$ as defined in the dataset and as shown in Figures 4.12, 4.13 and 4.14. The ground truth annotations are obtained[3] from the same authors [58, 59] Only people inside the Area of Interest AOI are considered. They are defined as:

- the annotations of the ground truth are enlarged by a 25 cm buffer in order to manage imprecision
- more than 50% of the ground truth area must be inside the AOI. Other people are thus not considered in the evaluation

Camera calibration and time synchronization errors are present in the dataset as specified in [28, 58]. We also encountered this problem as shown for example in Figure 4.8. We projected a cylinder of unit radius. It can be noticed that the projections do not converge to one point. For this dataset, the maximum number of people simultaneously monitored in the AOI is 8.

[3]The dataset and ground truth is available at: http://web.eee.sztaki.hu/~ucu/3dmpp/gt_citycenter.tar.gz.

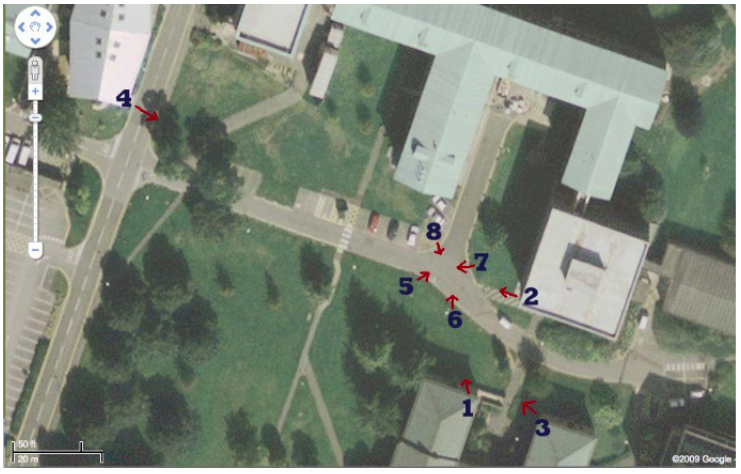


Figure 4.7: Location of PETS 2009 cameras on Google Maps [28]. Cameras 1, 2, and 3 are used.

Figures 4.9 and 4.10 show some frames across the three camera views of the PETS 2009 dataset. The dataset contains scene with full and partial occlusions, e.g. the person occluded by the post lighting (see Figure 4.9(a)), or the person barely noticeable behind the tree (see Figure 4.9(c)). The third view suffers from heavy clutter due to the tree. There is also significant color variation in the third view compared to the other two views (e.g. see Figures 4.9(a) and 4.9(c)).

4.4.2 Evaluation Measures

In order to compare our algorithm with the other ones, and to tune the parameters of our method, we need to quantitatively evaluate our results.

For this, we need a ground truth that can be available or manually obtained. Thus, let us assume that we know the real ground occupancy of people in the scene represented by a rectangle R_i covering the area of each individual person across the two legs. These rectangular areas are present in both the real world coordinate system and the image coordinate system of each camera view. Furthermore, let us assume that these ground truth rectangles are immune to the effects of the calibration or time synchronization errors.

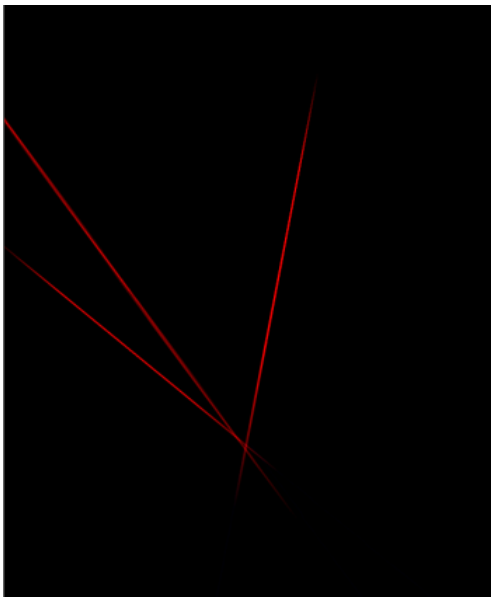


Figure 4.8: Illustration of a line-based kernel not converging to a singular point due to the calibration, time synchronization errors. Further details of the quantized impact of this error on PETS 2009 dataset can be seen in [78].

(a) View 1

(b) View 2

(c) View 3

Figure 4.9: Frame 683 from PETS 2009 dataset. Notice the occlusion introduced by the post lighting in the View 1. The proposed algorithm is designed to handle such partial occlusions. Notice also the heavy clutter for person behind the tree in View 3. Color variation across the camera views can also be seen.

(a) View 1

(b) View 2

(c) View 3

Figure 4.10: Frame 752 from PETS 2009 dataset. This particular frame shows cases of full and partial occlusion.

For numerical comparison, we define the concepts of matching the rectangles $\{R_1, R_2, \ldots, R_i, \ldots, R_M\}$ in the ground truth to the estimated position of pedestrians $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_j, Y_j), \ldots, (X_n, Y_n)\}$. Then, and as proposed by [58], a match function $m(i, j)$ is defined such as:

- $m(i, j) = 1$ if (X_j, Y_j) is inside the rectangle R_i.
- $m(i, j) = 0$ otherwise

Then, the Hungarian algorithm [79] is employed to find the best association and make the final matching $A = [a(i, j)]$ such as:

- $a(i, j) = 1$ if the detection i is associated to the jth rectangle of the ground truth.
- $a(i, j) = 0$ otherwise

From this matrix, Utasi et al. [58] define four measures:

- *The False Detections (FD).* This measure counts the number of detection without corresponding rectangle in the ground truth as it can be seen (see Figure 4.11(a)).
- *The Missed Detections (MD).* This measure counts the number of rectangle of the ground truth not assigned to a detection as illustrated Figure 4.11(b).
- *The Multiple Instances (MI).* This measure counts the number of times that several detection are assigned to the same rectangle of the ground truth (Figure 4.11(c)).
- *The total error (TE).* This measure is simply the sum of *FDs*, *MDs*, and *MIs*.

These measures allow the definition of the four parameters defined by [58] and used in this thesis for comparison as explained below.

(a) A False Detection represented by a cross and two good detection represented by circles

(b) Missed Detection: the rectangle of the ground truth is not associated to a circle

(c) Multiple Instances: a circle and a cross associated to the same rectangle

Figure 4.11: Example of detection errors. The rectangle represents the ground truth, the white circle represents a good detection the white cross represents a bad detection.

False Detections Rate (FDR)

The False Detections Rate (FDR) is the ratio of the detections not corresponding to a person to the total number of rectangles present in the ground truth:

$$FDR = \frac{FD}{M} = \frac{\#\{(X_j, Y_j) : \sum_{i=1}^{n} a(i,j) = 0\}}{M} \tag{4.4}$$

where M is the number of people really present and $\#F$ is the cardinal of the set F. It should be noted that the FDR may exceed 1.

Missed Detections Rate (MDR)

The Missed Detections Rate (MDR) is the ratio of the number of people which are not detected to the total number of rectangles present in the ground truth:

$$FDR = \frac{MD}{M} = \frac{\#\{R_i : \sum_{j=1}^{M} a(i,j) = 0\}}{M} \tag{4.5}$$

Here, $MDR \leq 1$.

Multiple Instances Rate (MIR)

The Multiple Instances Rate (MIR) is the ratio of number of people detected multiple times in a frame to the total number of rectangles present in the ground truth. Here, $MIR \leq 1$.

$$FDR = \frac{MI}{M} = \frac{\sum_{j=1}^{M} max(0, a(i,j) - 1)}{M} \tag{4.6}$$

Here, $MIR \leq 1$.

Total Error Rate (TER)

The Total Error Rate (TER) is the sum of FDR, MDR and MIR:

$$TER = FDR + MDR + MIR \tag{4.7}$$

Because FDR can exceed 1, therefore TER may also exceed 1.

It should be noted that the evaluation measure heavily penalises imprecision on the location of the person, i.e. it generates both a false detection and a missed detection[4].

[4]The evaluation code is available at: http://web.eee.sztaki.hu/~ucu/3dmpp/.

Table 4.1: Comparison of the proposed approach based on the Multi-camera Occupancy map Deconvolution (MOD) and an approach without pruning. The parameter sets are such that the TER is minimized for both method.

Sequence	**Method**	**TER**	**FDR**	**MDR**	**MIR**
PETS 2009	Without pruning	0.36	0.29	0.05	0.02
	MOD	0.13	0.03	0.10	0.00

4.4.3 Results

First, we present the experimental setup for our proposed algorithm. We model people with 175 cm high 3D cylinders as regularly done in literature [23, 58, 59]. The planar heights used to compute occupancy maps P_z are 211 planes, one plane for each centimeter between 0–210 cm, covering the range of possible human heights. The occupancy maps remain the same throughout this chapter and are generated at a 2 cm grid resolution.

In a first time, we compare in Table 4.1 our method with a simple detection based on a thresholding on the occupancy map. As expected, the number of false detection is considerably smaller. This proves that the proposed approach is able to manage the ghost phenomenon. The results on three images are presented in Figures 4.12, 4.13 and 4.14, where the rectangle is the bounding box in the ground truth and the white circle is the detection.

Furthermore, we have compared in Table 4.2 our method to five of the state-of-the-art techniques:

- Two methods proposed by Ren et al (Ren eq 6 and Ren eq 7). [14]
- Probabilistic Occupancy Map (POM) [23]
- 3D Marked Point Process (3DMPP) model [58, 59]
- Multiple Scene Plane Localization Method (MSPL) [13]

The results for POM and 3DMPP are reported from [58]. For MSPL, instead of a simple thresholding, we have extended the method in [13] to perform people detection by applying the same maxima detector to the occupancy maps.

We show that two methods proposed by Ren et al. [14] are unable to account for views with high variations of color. It is logical as both methods are based on color

(a) View 1

(b) View 2

(c) View 3

Figure 4.12: Results obtained on Frame 593 from PETS 2009 dataset. Notice the clutter introduced by the tree especially in View 3. The algorithm shows robustness to the color variation across the camera views.

(a) View 1

(b) View 2

(c) View 3

Figure 4.13: Results obtained on Frame 683 from PETS 2009 dataset. Notice that the occlusion introduced by the post lighting in the View 1 is handled.

(a) View 1

(b) View 2

(c) View 3

Figure 4.14: Results obtained on Frame 752 from PETS 2009 dataset. This particular frame shows cases of full and partial occlusion.

constancy. If we consider TER, it can be observed that we obtain improvements over MSPL and POM methods in PETS 2009. It remains close but does not surpass the performance compared to 3DMPP.

For the proposed MOD algorithm, Figure 4.15 shows TER plotted as a function of the τ and $\varnothing$ parameters. $\varnothing$ performs 3D reasoning whereas τ performs local analysis in the kernel space. The $\varnothing$ parameter fits to the specific radius of the pedestrians in the dataset (around 50 cm). The value of τ constraints two detections not to be too close. As in Figure 4.15, the best result for minimized TER is obtained for $\tau = 55$.

The process used to detect people is based on the principle of deconvolution. A true deconvolution with a space-varying kernel is not possible due to the computational constraints. This step is, therefore, approximated by a template similarity measure as presented in equation 4.3. This step is necessary as proved in Table 4.1 to avoid false detections due to ghost. However, this deconvolution step blurs the occupancy map and especially around the object-of-interest locations. For example, Figure 4.16 shows examples of the blurring effect observed across various images. This can also reflect towards the higher MDR rate in Table 4.2. This drawback led us to reconsider the method and propose a second method, presented in the following chapter.

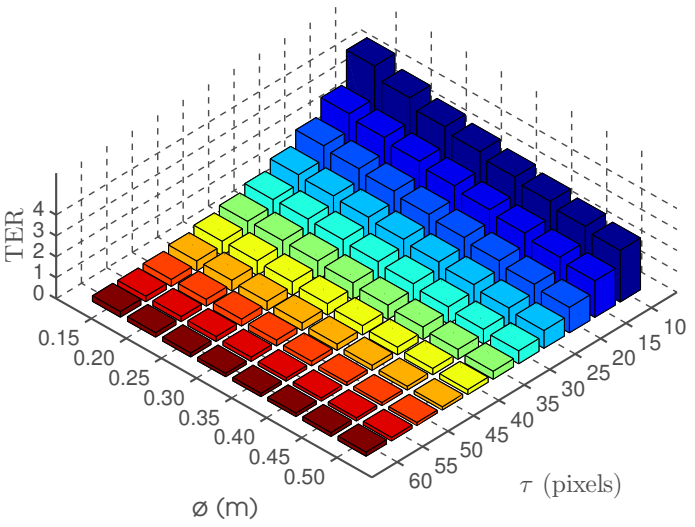


Figure 4.15: Evaluation of the MOD method with different parameter settings. Total Error Rate (TER) as a function of the τ and $\varnothing$ parameters for PETS 2009.

Table 4.2: Comparison of the proposed Multi-camera Occupancy map Deconvolution (MOD) method with the Multiple Scene Planes Localization (MSPL) method [13]. The parameter set is such that the TER is minimized. For MSPL: $\tau = 35$. For MOD and PETS 2009: $\varnothing = 0.50, \tau = 55$.

Sequence	**Method**	**TER**	**FDR**	**MDR**	**MIR**
PETS 2009	Ren eq6 [14]	0.88	0.20	0.68	0.00
	Ren eq7 [14]	0.89	0.20	0.69	0.00
	MSPL [13]	0.28	0.18	0.10	0.00
	POM [23]	0.25	0.18	0.07	0.00
	3DMPP [58]	0.12	0.02	0.10	0.00
	MOD	0.13	0.03	0.10	0.00

Like other projective methods relying on homographic occupancy constraints: the proposed algorithm is also influenced by the height of the cameras. We may study the two extreme cases. First, when cameras are at extremely low heights, then, the algorithm provides imprecise detections, yet precise height estimations. In the other extreme, when cameras are at the extreme top, the algorithm provides precise detections, but, imprecise height estimations. Thus, the results can further be improved with the increased height of the cameras or the introduction of another camera with increased height.

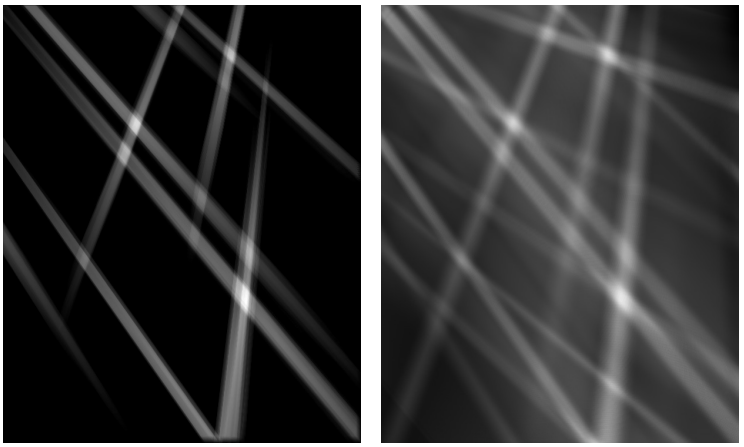
4.5 Summary

We have proposed a robust approach for performing people detection using multi-view reasoning in the multi-camera occupancy maps that avoids the problem of ghosts. We model the occupancy map by a convolution of 2D Dirac function and spatially varying kernel. This kernel depends upon the properties of an assumed 3D geometric primitive, the position and the camera parameters. Moreover, this allows to formalize the detection problem as a deconvolution that provides the object locations. We further introduce a novel parallelized approximation of the deconvolution, specific to our kernel responses.

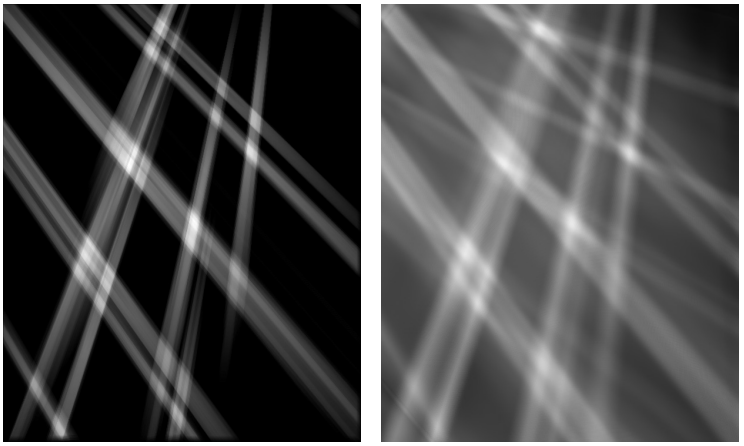
In terms of computational efficiency, the kernel generation step is time consuming but is performed offline, as the camera configuration does not change. For one diameter: the generation of 609×745 kernels using 211 planes takes approximate 5 days on a 64-core implementation. The computational times further increase with the diameter of cylinder; $50cm$ taking approximate 7 days. Similarly, for one

diameter, the estimated deconvolution step for 395 frames takes around 3 days to complete with our current implementation; however, this was performed with a single-core implementation. Combination of MATLAB, C++ and Java codes have been utilised. It is important to emphasise here that we have not focused on strict algorithmic efficiencies during the course of this thesis. Conversion of all the code into C++, memory optimizations (instead of disk writes), and a multi-core or GPU implementations (such as [31]) may result in real-time performances.

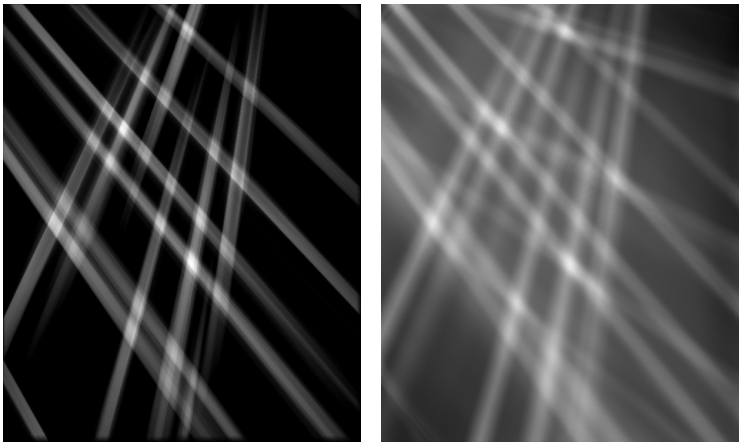
To improve the computational speed efficiency and to account for the blurring effect, we propose another idea in the following chapter. Instead of performing object modeling, we could first identify the relevant keypoints in the occupancy map. Once these points are identified, then we can use an idea similar to the template similarity measure in order to focus on ghost pruning. The identification of key locations in the beginning can lower the computational loads. We now introduce this proposition, based on the principle of hypothesis validation, in the next chapter.



(a) Occupancy Map (b) Template Similarity



(c) Occupancy Map (d) Template Similarity



(e) Occupancy Map (f) Template Similarity

Figure 4.16: Illustration of several images showing the blurring effect introduced in the occupancy maps due to the proposed approach.

Chapter 5

Ghost Pruning in the Multi-camera Occupancy Maps

5.1 Introduction

In this chapter, we present another method for multi-camera people detection based on the multi-view geometry and shape analysis performed on the occupancy maps. As in the previous chapters, we propose to create an occupancy map by the projection of foreground masks across all camera views on the ground plane and the planes parallel to the ground. This leads to significant values on locations where people are present, and also to a particular shape around these values. Moreover, a well-known ghost phenomenon appears i.e. when shapes corresponding to different persons are fused then false detections are generated. The first method proposed in this thesis models the process of occupancy map generation by a convolution of 2D Dirac functions and space varying kernel. People detection is then directly obtained with an approximation of deconvolution. This method improves detection results by avoiding ghosts but has two main drawbacks. (*i*) The deconvolution process involves a blurred effect on the occupancy map. (*ii*) The deconvolution is done for each pixel of the occupancy map with a kernel specific to this position, leading to important computation times. Thus, in this chapter, we begin with a robust detection of the candidate locations, namely keypoints, from the occupancy map based on a watershed transform. Then, in order to reduce the false positives, mainly due to the ghost phenomena, we check if the particular shape, for a person, is present or not. This shape, that is different

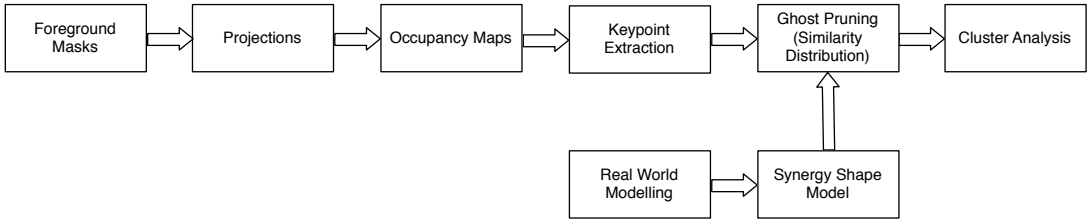


Figure 5.1: Block diagram for detection process by shaping the occupancy maps across the keypoints. Here synergy shape model refers to the proposed synthetic shapes.

for each location of the occupancy map, is synthesized for each keypoint, assuming the presence of a person, and with the knowledge of the scene geometry. Finally, the real shape and the synthetic one are compared using a similarity measure that is similar to correlation. Another improvement proposed in this chapter is the use of unsupervised clustering, performed on the measures obtained at all the keypoints. It allows to automatically find the optimal threshold on the measure, and thus to decide about people detection.

Figure 5.1 shows the block diagram for the proposed algorithm. As in the previous chapter, the proposed method makes use of the foreground masks obtained with the background subtraction method. The silhouettes are projected across multiple planes parallel to the ground plane. These projections are then merged to produce the occupancy map that has significant values at the locations corresponding to people or ghosts. Details on this part can be found in Chapter 3. We now proceed to the definition of keypoints in the following section[1].

5.2 Keypoint Extraction in the Occupancy Map

The first step consists in extracting locations of the occupancy map with important values denoted as keypoints. In order to be robust to noise and to impose a minimal distance between keypoints, the watershed based maxima selection algorithm (Algorithm 1) is employed. It leads to a set $\mathcal{P} = \{p_n\}, n = 1 \dots N$, of N keypoints that can correspond to people or ghosts. Actually, we know that ghosts may be detected as there are high probability locations other than the objects represented in the ground truth. Ghosts occur as a result of the intersection

[1] A part of this chapter appeared in the proceedings of the 10th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2015 [80].

of non-corresponding regions along the lines joining the camera centers and the object of interest. Therefore, a step is now necessary to distinguish between real people and ghosts.

5.3 Ghost Pruning with Shape Analysis

Some false positives or ghosts appear in the set of keypoints obtained (see Figure 3.12). Evans *et al.* [24] found that the ghost detections are probable along the lines from the camera centre to the centre of the objects of interest, intuited as "star" shape aound the object location, having "streaked legs" corresponding to the lines (see Figure 3.12). In this work, we propose a novel model that plays a role in ghost pruning using the shape cues defined around these "star" shapes and "streaked legs". Liu *et al.* [81], for robust auto-calibration, models the pedestrian blobs using two end points of the axes of the ellipses, represented by the vanishing point and estimating the 3D blob heights resembling the real world distribution of human heights. Following this, we define a shape for each person, represented by the longitudinal axis of a cylinder in the 3D coordinate system of the scene as represented in Figure 5.2(b).

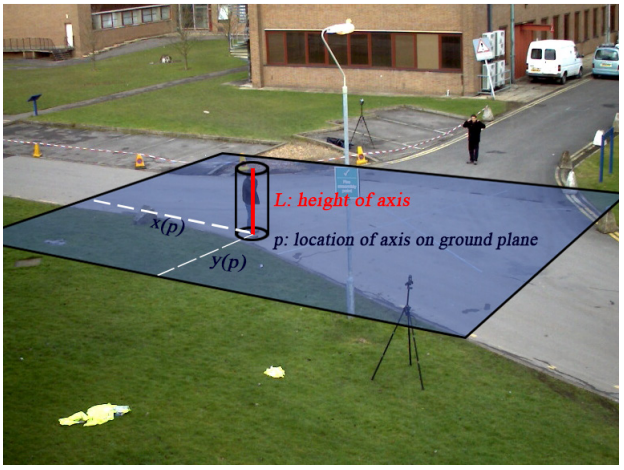
Let us assume that the people are standing on a flat ground. We monitor a rectangular Area of Interest (AOI) in the P_0 ground plane, and we attempt to model the shape of each possible pedestrian in $\mathcal{P}$. Thus, the free parameters of the given longitudinal axis of the cylinder are its coordinates $\mathbf{p} = (x, y)$ in the ground plane and its length L. This is illustrated in Figure 5.2(b).

We employ a discrete space of objects in the ground plane of the AOI, consisting of $S_W \times S_H$ locations. For each keypoint p_n detected in this space, synthesized camera views $\mathcal{I}_{n,v}$ are generated using the camera calibration matrices. $\mathcal{I}_{n,v}$ corresponds to the projection of the longitudinal axis situated at p_n in the camera view v . Let $\mathcal{I}_n$ denotes the set of synthetic images created for the keypoint p_n:

$$\mathcal{I}_n = \{\mathcal{I}_{n,v}\}\,,\ v = 1 \ldots V. \tag{5.1}$$

All these images are projected on the ground plane P_0 and several planes P_z parallel to the ground planes with different heights z. Let us denote $\mathcal{P}_{n,v,P_z}$ the

(a) Camera View 1



(b) Synthetic View

Figure 5.2:
(a) Original image
(b) Longitudinal axis (in red) of a 3D cylinder modeling the person. The height of this axis is denoted by L.

projection of $\mathcal{I}_{n,v}$ on the plane P_z. All these projections are fused, as detailed in equation 3.3, to produce a synthetic occupancy map $\mathcal{SO}_n(x, y)$.

The goal here is to use the fact that a person generates not only a significant value in the occupancy map but also a particular shape around this point — a shape that is not present in case of a ghost. Illustration of synthetic occupancy maps $\mathcal{SO}_n(x, y)$ are presented in Figure 5.4 where the synthesized models [see Figures 5.4(a)-(c)] have more similarity with the real occupancy map, as compared to the ghosts [see Figures 5.4(d)-(f)]. In visual terms, we may consider these synthesized occupancy maps as asterisk. However, for ghosts, streaks of the asterisk in

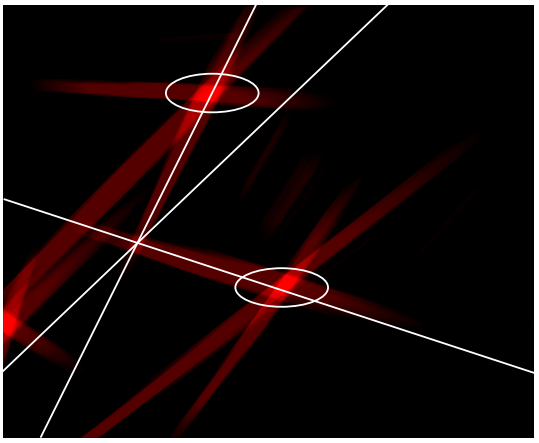


Figure 5.3: Justification of the intersection measure between forms: when several people are present, a simple correlation measure is perturbed by the legs of the star shape produced by the other persons (white ellipse).

the real occupancy map are inconsistent with the streaks of the synthesized occupancy map patterns, pointing towards a probable anomalous detection. Thus, it seems that ghost removal can be achieved by comparing real and ideal asterisks.

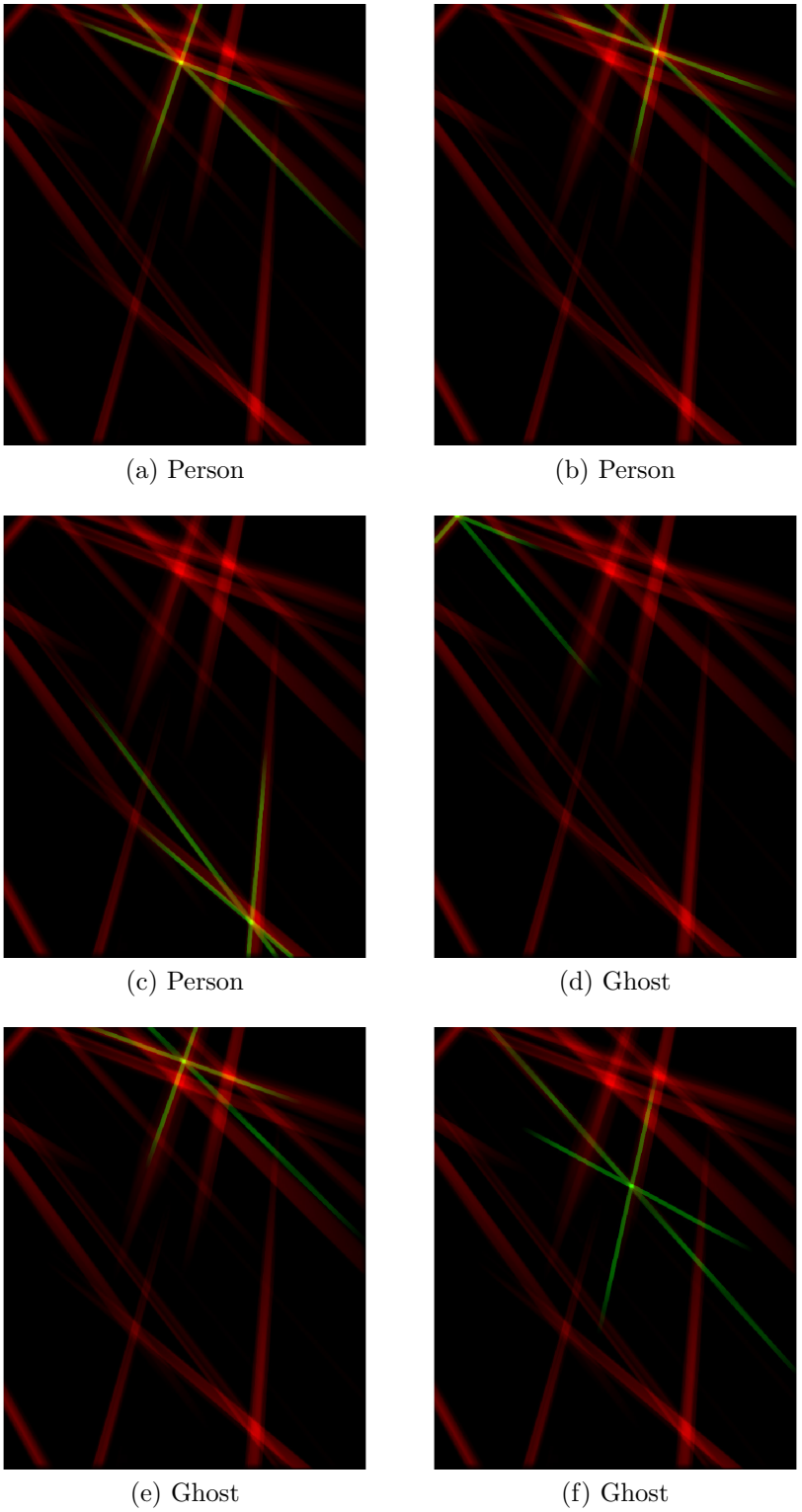
Therefore, for each keypoint p_n, we generate a synthetic occupancy map $\mathcal{SO}_n(x, y)$, that corresponds to the occupancy map to be observed if a person is present at the location p_n: the person is modeled by a vertical line, the axis of a cylinder, with height L (see Figure 5.2). By comparing $\mathcal{SO}_n(x, y)$ to the real occupancy map $\mathcal{O}(x, y)$, we can define a similarity measure $\hat{\mathbf{D}}_\mathbf{n}$ associated to the keypoint p_n between the shapes. Two similarity measures have been proposed. The first one, is the simple 2D correlation defined as:

$$\hat{\mathbf{D}}_\mathbf{n} = \frac{\sum_x \sum_y \mathcal{O}(x, y)\mathcal{SO}_n(x, y)}{\sum_x \sum_y \mathcal{SO}_n(x, y)} \tag{5.2}$$

The second measure looks like an intersection between the two shapes and is inspired from histogram intersection [74]:

$$\hat{\mathbf{D}}_\mathbf{n} = \frac{\sum_x \sum_y min(\mathcal{O}(x, y), \mathcal{SO}_n(x, y))}{\sum_x \sum_y \mathcal{O}(x, y)} \tag{5.3}$$

This last measure is justified when several people are present in the scene (Figure 5.3). Indeed, a simple correlation measure is affected by important values of the occupancy map induced by the legs of the star shape produced by these people.



(a) Person (b) Person

(c) Person (d) Ghost

(e) Ghost (f) Ghost

Figure 5.4: Illustration of the proposed method on frame 593 of the PETS 2009 dataset. The Occupancy Shape Model is represented by the green color whereas the occupancy maps are red. It can be observed that the overlap between both is higher for the persons (a),(b),(c) versus the ghosts (d),(e),(f).

These two measures $\hat{\mathbf{D}}_\mathbf{n}$ can be understood as the confidence of the hypothesis that a person is present at the keypoint p_n.

5.4 Detection based on thresholding

A threshold must be estimated on $\hat{\mathbf{D}}_\mathbf{n}$ to decide if the keypoint p_n corresponds to a real person or a ghost. Rather than to optimize it according to the ground truth, we decided to setup a step to automatically fix the threshold. Given a particular cameras configuration, this allows to obtain an autonomous system that adapts itself to each application. So, given the similarity measures distribution over a dataset, we can proceed towards univariate cluster analysis to group the similarity measures into class intervals corresponding respectively to people and to ghosts.

Cluster analysis is an unsupervised learning technique which uses a set of unlabeled data as input and tries to determine an intrinsic grouping in the set. Due to cluster analysis, the decision threshold for a keypoint p_n, to be a person or a ghost, is automatically computed. Most clustering or vector quantization algorithms can be classified into partitional or hierarchical algorithms [82] (see Figure 5.5). Hierarchical clustering algorithms do not require pre-specification of the number of clusters, are primarily deterministic, but computationally expensive. Partitional or flat clustering algorithms define a set of disjoint clusters and are suited for large datasets where computational efficiency is important. However, as no consensus is present on this issue [83], therefore, we use both partitional and hierarchical methods.

For **hierarchical clustering**, we use the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) agglomerative clustering method [84] with an Euclidean distance for the generation of the distance matrix. The UPGMA algorithm constructs a rooted tree (dendrogram) that reflects the structure present in a pairwise similarity matrix (or a dissimilarity matrix). At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters A and B is taken to be the average of all distances between pairs of objects "x" in A and "y" in B, that is, the mean distance between the elements of each cluster. We select UPGMA clustering because it provides a suitable trade-off between the complete-link method's sensitivity to outliers, and, single-link method's sensitivity to form dendogram chains longer than the intuitive notion of compact

and spherical clusters. The work in [85] also selects agglomerative clustering technique for human detection in 3D space.

For **partitional clustering**, we use two methods:

- the univariate Kernel Density Estimation (KDE) with Epanechnikov kernel [86] and local minimum to separate the clusters.
- the Mixture of Gaussians Expectation Maximization (MoG-EM) method (univariate, unequal variance) [87]. In this case, we assume as a priori that the number of real objects exceed ghosts, hence we can have a distribution for ghosts with lower variance, centered around a mean corresponding to the lower measure of similarity.

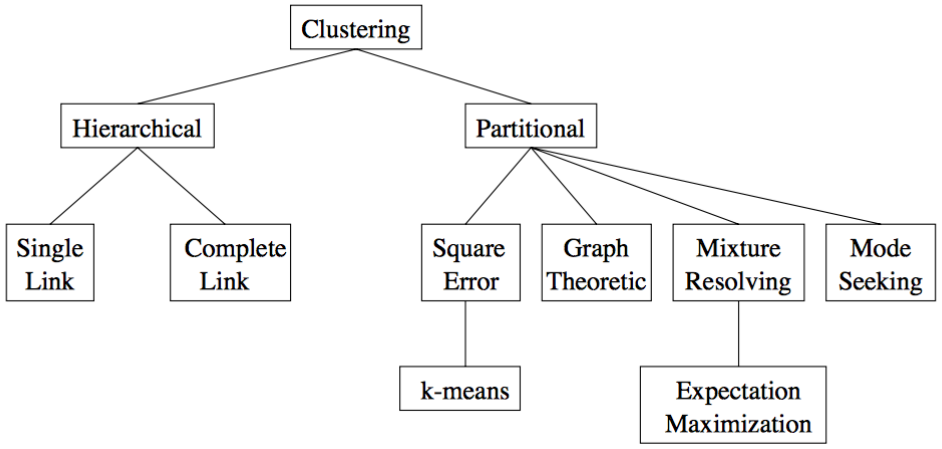


Figure 5.5: Taxonomy of data clustering techniques. Reproduced from [82].

Algorithm 3 People detection by ghost pruning in multi-camera occupancy maps

for all X of the occupancy map **do**
  for all Y of the occupancy map **do**
    compute the synthetic occupancy map $\mathcal{SO}_n(x, y)$ (parameter: L)
  end for
end for
for all time k **do**
  use the watershed algorithm to extract keypoints p_n (parameter: τ)
  for all the keypoints **do**
    compute the similarity measure $\hat{\mathbf{D}}_\mathbf{n}$ using Equation 5.3
  end for
end for
automated thresholding for detection (univariate unsupervised clustering)

Table 5.1: Comparison of the different similarity measures. The best parameters for minimum *TER* are used. Parameters: $\tau = 31$ pixels, $L = 175$cm.

Method	**TER**	**FDR**	**MDR**	**MIR**
Intersection Eq. (5.3)	0.076	0.025	0.051	0.000
Correlation Eq. (5.2)	0.096	0.043	0.053	0.000

Hierarchical clustering is more suited to our univariate data because it doesn't have enough structure, relative to multi-dimensional data, and the computational costs are not important. Moreover, KDE requires a bandwidth specification, while a prior has to be defined in the case of MoG-EM method. For all three methods, we suppose *a priori* that the data are divided into two classes or clusters. Results of the three algorithms are presented in the following section.

5.5 Results on PETS 2009

The method presented in this chapter is denoted as OSM, i.e. Occupancy Shape Matching, in the following. As in Chapter 4, we present first the experimental setup. The foreground masks are generated using the default parameters as defined in [71]. Multi-planar projections are generated at a constant 2 cm resolution as proposed in [71]. For similarity with [23, 58, 59], we fix the height L to 175 cm, and the occupancy maps are generated for 56 different planes between 155 and 210 cm. The proposed method has thus only one free parameters: the tolerance threshold τ. This is an important advantage compared to other methods that have several parameters [58, 59].

5.5.1 Study of the similarity measure

We present two ways to compare real occupancy map and synthetic one (see Equations 5.2 and 5.3). In this section, we evaluate these two measures. Evaluation results are presented in Table 5.1.

The best results are obtained for computations based on the similarity measure that uses shape intersection. Therefore, we focus only on this similarity measure in the following.

5.5.2 Influence of the parameter τ

The proposed method has only one parameter: the parameter τ used to detect keypoints with the watershed algorithm. Moreover, we fix the height of people to $L = 175$ cm as often done in literature [23, 58, 59] (even if all persons do not measure 175 cm) and we study the influence of both parameters on the performance of the detection. Figure 5.6(a) shows Total Error Rate (TER) plotted as a function of τ and L parameters. We observe that the TER depends mainly on τ parameter. Higher values of τ tend to merge all the keypoints, and the lower values tend to introduce multiple keypoints for the same "star" shape. The parameter L seems to have no influence on the TER. We also draw the Precision/Recall and ROC curves for different L values in Figure 5.6(b). These last curves have been obtained by fixing τ to its optimal value ($\tau = 31$ pixels) found in Figure 5.6(a) and varying the detection threshold. The curves show a small improvement when L increases. In the following, we keep L to 175 cm for three reasons. (*i*) This value is commonly used in literature. (*ii*) This is the average height of human. (*iii*) This parameter seems to have no influence when using the detection based on an automatic threshold selection.

5.5.3 Influence of the methods used to select the optimal threshold

In section 5.4, we present three clustering techniques to select the optimal threshold for detection. It can be observed from Table 5.3 that the results remain consistent in spite of the clustering technique used. So, in the following, we use the UPGMA method that is the simplest to use.

Table 5.2: Comparison of proposed OSM with other techniques using the best parameters to minimize *TER*. Parameters: $\tau = 31$ pixels, $L = 175$ cm.

Sequence	**Method**	**TER**	**FDR**	**MDR**	**MIR**
PETS 2009	MSPL [13]	0.28	0.18	0.10	0.00
	POM [23]	0.25	0.18	0.07	0.00
	3DMPP [58]	0.12	0.02	0.10	0.00
	OSM	0.08	0.03	0.05	0.00

5.5.4 Comparison with other works

We have compared our method to three others of the state-of-the-art techniques:

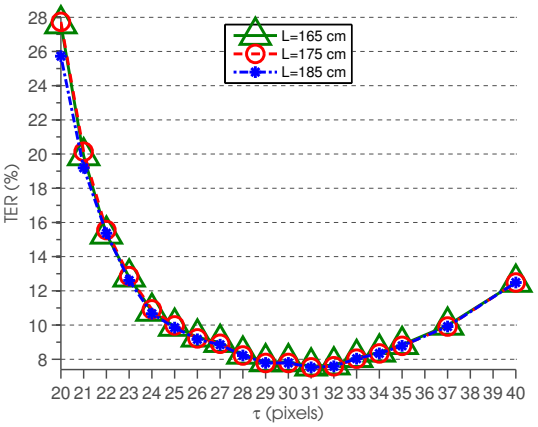
- POM [23]
- 3DMPP [58]
- MSPL [13]

The results of the four algorithms are reported in Table 5.2. Considering TER, we observe a 4% improvement versus 3DMPP, 17% versus POM, and 20% versus MSPL. All the algorithms have a negligible Multiple Instances Rates (MIR), approximately zero. However, the proposed approach presents now a really small False Detection Rate (0.03) combined with a small Missed Detection Rate (0.05). Let us recall that without ghost removal, and thus by considering all the keypoints detected by the watershed algorithm as people, the False Detection Rate was 0.291 (Table 4.1). This proves that the last step, which compares real occupancy map and the synthetic ones, is effective in removing the ghosts.

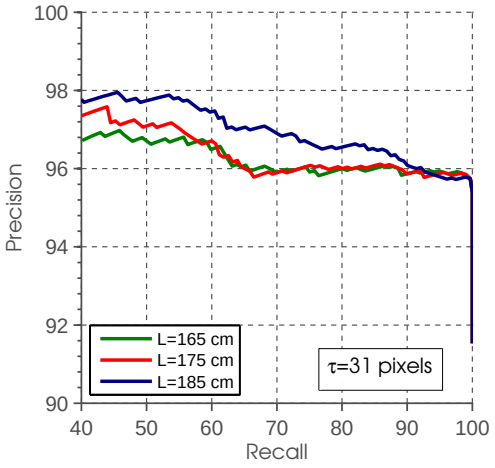
Figures 5.7, 5.8, and 5.9 show the successful detection obtained with our algorithm in challenging scenarios. Figure 5.7 shows the phenomenon of people outside the AOI affecting the results. In this case, the detection outside the AOI can simply be pruned with the location priors. However, our algorithm also accounts for extended erroneous detection inside the AOI. Figure 5.10 shows an example of a frame with missed detection.

Table 5.3: Comparison of the different clustering techniques. The best parameters for minimum *TER* are used. Parameters: $\tau = 31$ pixels, $L = 175$ cm, and for KDE, the bandwidth of the kernel is 0.04.

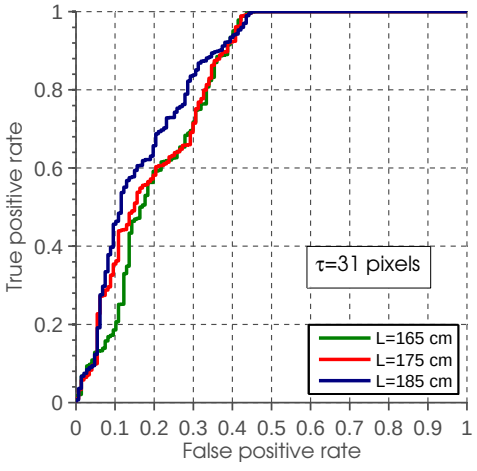
Method	**TER**	**FDR**	**MDR**	**MIR**
KDE	0.077	0.026	0.051	0.000
MoG-EM	0.076	0.025	0.051	0.000
UPGMA	0.076	0.025	0.051	0.000



(a)



(b)



(c)

Figure 5.6: Evaluation of the proposed Occupancy Shape Model with different parameter settings. (a) Total Error Rate (TER) as a function of the τ and L parameters. (b) Precision/Recall curves and (c) Receiver Operating Characteristic (ROC) curves in function of L.

(a) View 1

(b) View 2

(c) View 3

Figure 5.7: Results obtained on frame 593 of the PETS 2009 dataset. The clutter and color variation is accounted for. We show a detection outside the Area of Interest (AOI). These detections occur due to the people present outside the AOI. In this particular case, they are simply avoided as being outside the AOI. However, our method also shows robustness to the phenomena which occur inside our space of interest.

(a) View 1

(b) View 2

(c) View 3

Figure 5.8: Results obtained on frame 752 of PETS 2009 dataset. All people present inside the Area of Interest are detected accurately despite of the full and partial occlusions.

(a) View 1

(b) View 2

(c) View 3

Figure 5.9: Results obtained on frame 784 of PETS 2009 dataset. All people present inside the Area of Interest (AOI) are detected accurately. Notice the case of a person at extreme left of the View 2. He is not inside AOI of this camera but in the others. The detection is correctly estimated.

5.6 Summary

In this chapter, we have presented another technique for a multi-camera system to robustly detect people using the knowledge of the scene geometry. We begin with the creation of the occupancy maps by merging all the projected views on the ground plane and the planes parallel to it. The moving objects not only produce significant values in the occupancy map but also a particular shape around them. This chapter proposes a solution to perform people detection and to avoid ghosts: at each candidate detection, we verify if the particular shape for an ideal person is present. This idea has been implemented, and we focus on the three tasks in this chapter: (i) how to find the points corresponding to potential candidates, (ii) the creation of the Occupancy Shape Model, and (iii) which tolerance can be accepted when studying the shape around the candidate detection.

In chapter 3, we presented the popular solution that consists to threshold the occupancy. This method presents a main drawback: ghosts are detected at locations where several shapes induced by different people overlap. We further utilized the idea of people detection using 3D geometric primitive shapes in chapter 4. However, the concept suffered from blurring effects and has limited computational performance. These effects, of blurring and reduced computational performances, are overcome in the particular technique presented in this chapter. As a final improvement, we have also presented the application of unsupervised learning techniques in order to automatically compute the decision threshold.

In terms of computational efficiency, the shape model takes 3 days to compute with a 64-core implementation. This computation is performed only once provided the calibration remains constant. Moreover, the similarity matching also takes significantly less time i.e. around 15-20 minutes on a single-core implementation for 395 frames. Combination of MATLAB, C++ and Java codes have been utilised. As mentioned in the previous chapter, it is important to emphasise here that we have not focused on strict algorithmic efficiencies during the course of this thesis. Conversion of all the codes into C++, memory optimizations (instead of disk writes), and a multi-core or GPU implementations (such as [31]) will result in real-time online performances.

In the following chapter, we compare the proposed methods on two other publicly available datasets. We explain the datasets and the experimental setup. Finally,

we proceed with the presentation of the quantitative comparisons and analysis of our results.

(a) View 1

(b) View 2

(c) View 3

Figure 5.10: Results obtained on frame 683 of PETS 2009 dataset. There is one missed detection in our estimations for this scene.

Chapter 6

Experiments on Other Datasets

6.1 Introduction

We have presented two different novel approaches for performing people detection earlier in this thesis. We have shown the efficiency of our approach on the challenging public dataset PETS 2009. In this chapter, we extend the validity of our approaches by testing it on other datasets as well. For this purpose, we have selected two more datasets. Both of these datasets are public, challenging, widely-used and established in the community of visual surveillance. In a first step, we introduce these datasets. Following this, we present the experimental setup. The evaluation strategy is similar to the last two chapters and as in [58, 59]. Finally, we conclude this chapter by presenting a summary of all the results obtained in the thesis [1].

6.2 Datasets

We have selected two more datasets from the *Performance Evaluation of Tracking and Surveillance* series of international workshops. The workshop has been running for over 10 years and addresses the scenario of multi-sensor visual surveillance in order to protect the critical infrastructures. We have already shown the efficiency of our approach on PETS 2009 [28]. We further proceed by testing it on

[1] This chapter appears as a section of a publication accepted in the proceedings of the 12th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), 2015.

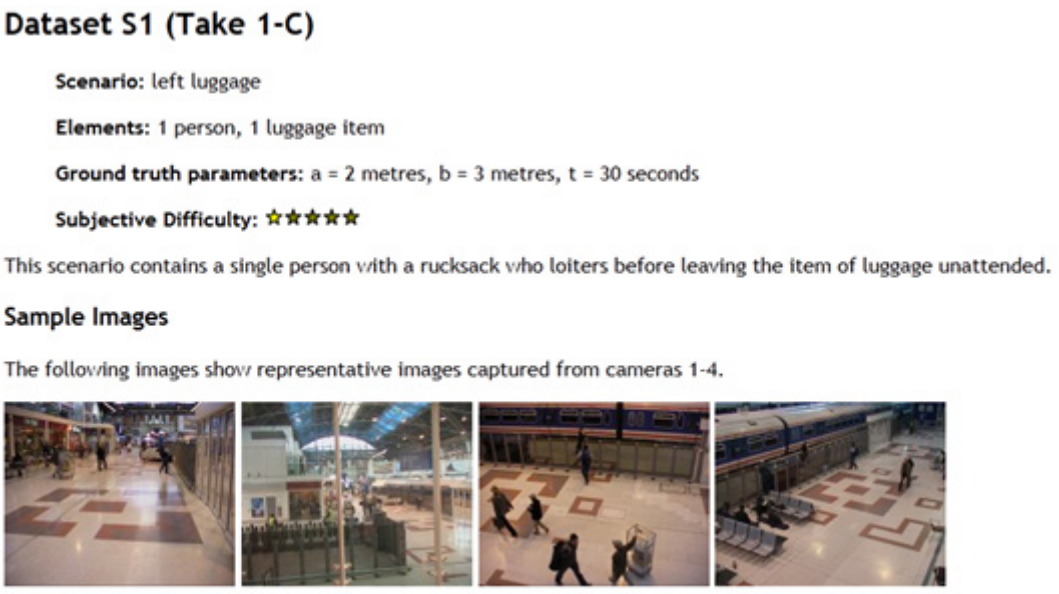


Figure 6.1: Sequence selection for PETS 2006: details as presented on the website of the workshop. Reproduced from [26].

PETS 2006 [26] and PETS 2007 [27] datasets. These datasets are gathered in a train station and at an airport respectively. The selection of these two datasets was of a particular interest to the goals of *French institute of science and technology for transport, development and networks* (IFSTTAR) i.e. towards the application of this thesis in the transportation scenarios.

6.2.1 PETS 2006

The second dataset which we used in this thesis is the PETS 2006 dataset [26]. This dataset is a multi-sensor sequence captured in a real-world, indoor scenario of train station. The aim of this specific dataset is to perform luggage detection, more specifically generate automated alerts for left luggage scenarios. Figure 6.1 shows description of one such sequence as captured from the PETS 2006 website [26]. There are four camera views capturing the scene for a left luggage scenario. The ground truth - provided by the dataset creators - has been made for the luggage detection scenario only. Therefore, we have to propose adjustments in order to account for multi-camera people detection. For this purpose, we use the same guidelines as earlier in the thesis, and as proposed by [58, 59].

We have used the last two cameras of the S1 sequence. The selection of cameras is based on the wide-base line criteria. The motivation is to produce an AOI

which covers relatively varying positions and topologies across the camera views. This selection of camera views and of the AOI can be seen in Figure 6.2 where views have been labeled as View 1 and Views 2. We manually annotated people presence in 120 non-consecutive frames per camera, across the two views, in the sequence. There are total 3021 frames in the S1 sequence. The camera calibration information using the Tsai framework [43] is provided and used. In case the calibrations are not available then the camera homography can be calculated and used as in [13]. The maximum number of people visible anywhere in the scene is 12.

Figure 6.3 shows a random example from our dataset. This dataset suffers from partial occlusions, presence of objects other than humans (bags, trolleys), texture-less surfaces such as the glass partitions installed next to the train (see Figure 6.3), reflections particularly in View 1 (see Figure 6.3(a)). The selection of AOI also introduces constraints of perspective selection. If a person is inside the AOI can produce artifacts in the occupancy maps (shown later in this chapter, see Figure 6.13). The reflections and the texture-less surfaces introduce noise and errors in the foreground mask extraction process (see Figure 6.4) that can induce over-detections in the foreground mask. Compared to PETS 2009, the foregrounds masks extraction is more difficult and leads to several errors (see Section 3.2.1). We keep the similar background subtraction method throughout the thesis. Nevertheless, we did extensively test several methods with the BGS library [88], none of which could give significantly better foreground masks. For a similar problem, Kiss et al. [89] propose post-processing of the foreground masks, however we do not perform such procedure to gauge the efficiency of our method while facing the noise of foreground masks.

6.2.2 PETS 2007

We use another public dataset from the transportation scenario, the PETS 2007 dataset [27]. This dataset is also a multi-sensor sequence and is designed for benchmarking challenges related to loitering, theft, and abandoned luggage. This scene is a combination of both indoor and outdoor scenarios, including sunlight variations across the sequence. The dataset was captured inside the departure area of an airport. The details of the specific random sequence we use, as available on the website of PETS 2007, are shown in the Figure 6.5.

For PETS 2007, we selected the cameras 2 and 3 of the S8 sequence. Similar to other sequences, the AOI is defined such that it is visible from all cameras (see Figure 6.6). The cameras above are selected to cover these AOI from relatively varying positions and topologies. The maximum number of people visible anywhere in the scene at a given time is 17. Thus, PETS 2007 has a higher density of pedestrians compared to PETS 2006.

We manually annotated a total of 120 non-consecutive frames for PETS 2007. There are total 3000 frames in the S8 sequence. Figure 6.7 shows an example of one such frame. We use the camera calibration data which is also available for this dataset. The tools and criteria as defined in [58] are used for annotation purposes. We notice presence of errors in the foreground masks (see Figure 6.8), errors of camera calibration (see the slight variation of the estimated position in Figure 6.13(c) and Figure 6.13(d)), and the perspective effect due to the AOIs (see Figure 6.13).

6.3 Experimental Setup

For both algorithms, foreground masks have been estimated using the default parameters of the process defined in [71] (see Chapter 2). The foreground masks contain noise and errors (see Figures 6.4 and 6.8). Three methods are compared in this section:

Figure 6.2: Area of Interest (AOI) selection for PETS 2006 in two views. We have selected the last two cameras from the original sequence. We call the left one View 1, and the other View 2. We select these two views as they are wide-base line stationary cameras.

(a) PETS 2006 View 1

(b) PETS 2006 View 2

Figure 6.3: Example of a frame in PETS 2006.

- The **Multi-camera Occupancy map Deconvolution (MOD)** approach introduced in Chapter 4.
- The **Occupancy Shape Model (OSM)** approach introduced in Chapter 5
- The **Multiple Scene Planes Localization (MSPL)** method [13] that is the most popular in literature.

For all methods, the planar heights used are 211 planes, one plane for each centimeter between 0–210 cm, covering the range of possible human heights and we fixed the height of people to 175 cm. The occupancy maps continued to be generated

(a) PETS 2006 View 1

(b) PETS 2006 View 2

Figure 6.4: Example of foreground masks extracted from the frames presented in the Figure 6.3. Notice the errors in the foreground masks that account for two people respectively outside the AOI and behind the glass.

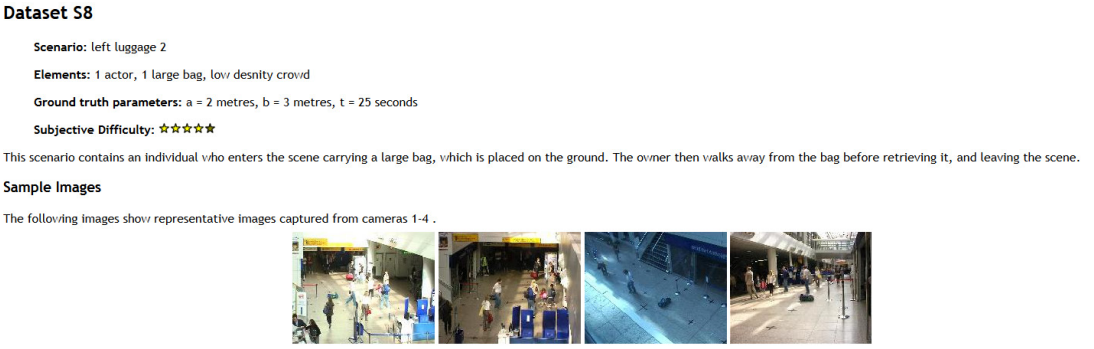


Figure 6.5: Sequence selection for PETS 2007: details as presented on the website of the workshop. Reproduced from [27].

at a 2 cm grid resolution using the multi-planar, multi-camera projections. Moreover, we use the similarity measure based on the shape intersection (equation 5.3). This is because, as we proved earlier, it is more efficient for our application (section 5.5.1). The evaluation measures used are the same as defined earlier and defined in section 4.4.2.

For the **Multi-camera Occupancy map Deconvolution (MOD)** approach introduced in Chapter 4, we report optimal results obtained to minimize the Total Error Rate (TER). So both parameters (diameter of the cylinder $\varnothing$ and tolerance threshold τ) have been optimized. We also report the curves showing the influence of these parameters.

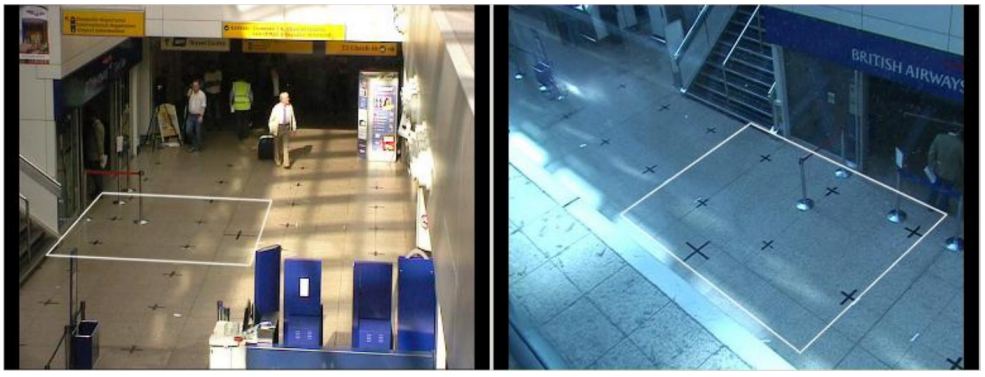
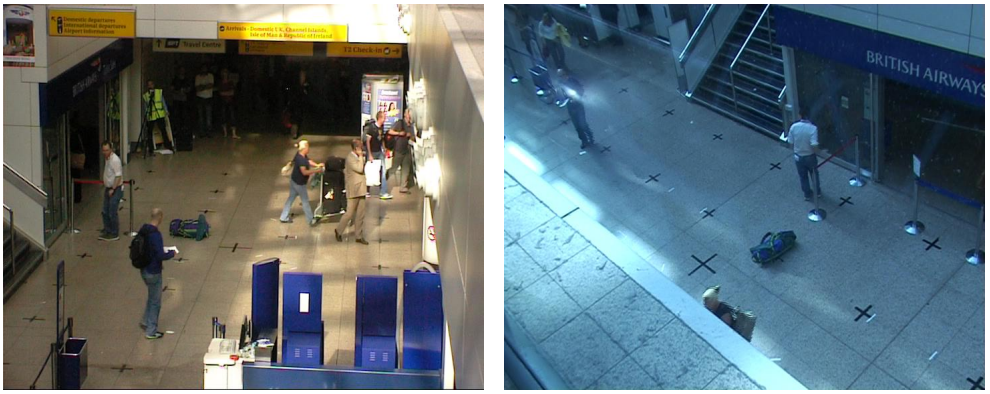


Figure 6.6: Area of Interest (AOI) selection for PETS 2007. We have selected the last two cameras from the original sequence. We call the left one View 1, and the other View 2. We select these two views as they are wide-based line stationary cameras.



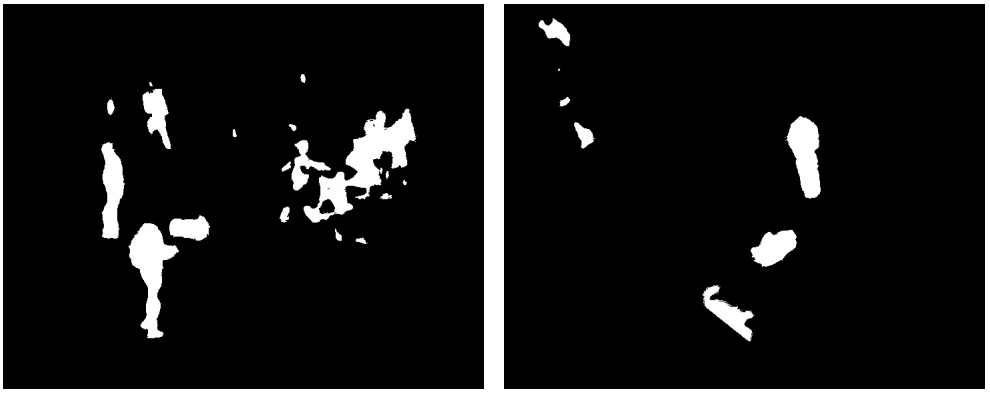
(a) PETS 2007 View 1

(b) PETS 2007 View 2

Figure 6.7: Example of a frame in PETS 2007.

For the **Occupancy Shape Model (OSM)** approach introduced in Chapter 5, results have been optimized according to only one parameter: the tolerance threshold τ that has been optimized in order to minimize the TER. The decision threshold is obtained by choosing the optimal operating point on the Receive Operating Characteristic (ROC) Curve [90].

For the **The Multiple Scene Planes Localization (MSP)** method [13], people detection has been achieved using the proposed watershed transform (see Algorithm 1) in order to made a fair comparison.



(a) PETS 2007 View 1

(b) PETS 2007 View 2

Figure 6.8: Example of foreground masks on a frame in PETS 2007 dataset. Notice the errors in the foreground masks.

Table 6.1: Comparison of the proposed Multi-camera Occupancy map Deconvolution (MOD), Occupancy Shape Model (OSM) methods with the Multiple Scene Planes Localization (MSPL) method [13]. The parameter set is such that the TER is minimized. For MSPL: $\tau = 35$. For MOD: $\varnothing = 0.50, \tau = 35$. For OSM 6: $\tau = 30$.

Sequence	Method	TER	FDR	MDR	MIR
PETS 2006	MOD	0.10	0.00	0.10	0.00
	OSM	0.12	0.00	0.12	0.00
	MSPL [13]	0.28	0.18	0.10	0.00

6.4 Results

We now present the results obtained on PETS 2006 and PETS 2007 datasets, using MSPL, MOD and OSM techniques. Finally, we present a brief summary of the results and comparisons obtained through the course of this thesis.

6.4.1 PETS 2006

We report the evaluation results obtained from the three algorithms in Table 6.1. The best parameters at which these results are obtained are also mentioned with the table. Considering the Total Error Rate (TER), we observe an approximately 18% improvement versus MSPL. Overall, we obtain comparable results for MOD

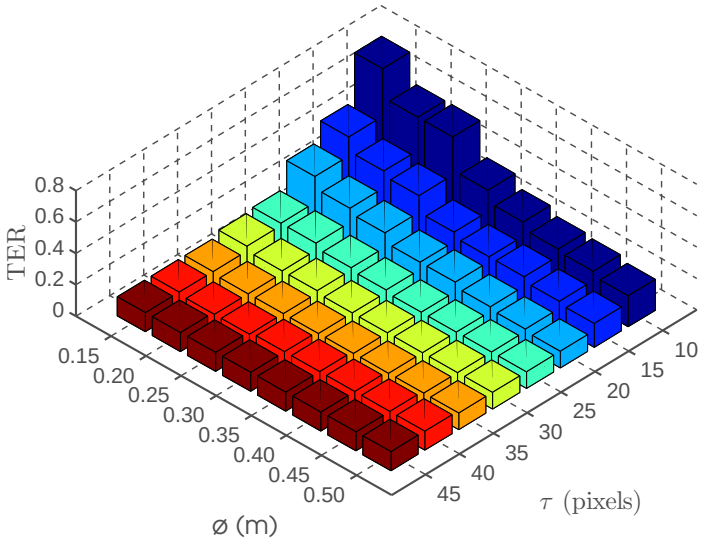


Figure 6.9: Evaluation of the MOD method with different parameter settings. Total Error Rate (TER) as a function of the τ and $\varnothing$ parameters for PETS 2006.

and OSM methods. We can notice that both methods account in successful removal of the false detection due to ghosts as the FDR (False Detection Rate) is really smaller than that of MSPL. However, OSM increases the Missed Detection Rate (MDR) negligibly while removing the FDR. We can see an example of such a case in Figure 6.10. This is an isolated case for PETS 2006 in which the detection, while identified as a keypoint initially, is removed later by the similarity assignment with the corresponding shape model. The efficiency of OSM could be affected at the border regions, as the star-shaped structure is rather incomplete. Increasing the heights of the cameras, introducing better normalization at border region could help in such isolated scenarios. Furthermore, as the proximity between people in PETS 2006 is not important, or in other words, as there is no situation tending toward crowded scenarios, one may notice that even the MSPL method performs relatively fine on PETS 2006. Later, we see a different scenario in the PETS 2007 dataset.

For the MOD method, we also present a curve in Figure 6.9 that shows TER plotted as a function of the τ and $\varnothing$ parameters. We remind that the $\varnothing$ performs 3D reasoning based on the 3D primitive shape, fitting a cylinder of particular radius in 3D coordinates. On the other hand, τ performs local analysis during the extraction of maxima using the watershed algorithm. Similar to PETS 2009, optimal values for $\varnothing$ resemble those found in [31].

6.4.2 PETS 2007

Table 6.2: Comparison of the proposed Multicamera Occupancy map Deconvolution (MOD), Occupancy Shape Model (OSM) methods with the Multiple Scene Planes Localization (MSPL) method [13]. The parameter set is such that the TER is minimized. For MSPL: $\tau = 35$. For MOD and PETS 2007: $\varnothing = 0.40, \tau = 40$. For OSM and PETS 2007: $\tau = 20$.

Sequence	**Method**	**TER**	**FDR**	**MDR**	**MIR**
PETS 2007	MOD	0.36	0.08	0.28	0.00
	OSM	0.17	0.00	0.17	0.00
	MSPL [13]	1.08	0.89	0.19	0.00

The evaluation results, with their best parameters, for PETS 2009 are reported in the Table 6.2. Considering Total Error Rate (TER), we observe an approximately 91% improvement with OSM versus MSPL and 72% between MOD and MSPL.

(a) PETS 2006
MOD

(b) PETS 2006
OSM

Figure 6.10: Example of a same frame being correctly processed by MOD but not OSM. The reason for this has been explained in the text.

(a) Result on PETS 2006 with MOD- view1

(b) Result on PETS 2006 with MOD- view2

Figure 6.11: Example of the estimated pedestrian locations in the PETS 2006 dataset. The MOD algorithm correctly distinguishes between the pedestrian and trolley and manages ambiguities of presence in Area of Interest (AOI) across views.

PETS 2007 dataset has dense situations and strong perspective effects. However, the proposed methods significantly improve the detection scores compared to MSPL.

OSM also performs notably better than MOD. The reason for this lies with the blurring effect that has already been explained in Chapter 4. Furthermore, we see that OSM has successfully removed false detection (False Detection Rate, FDR) without penalizing the missed detection (MDR). On the opposite, MOD has heavily penalized the missed detection to decrease the false detection - a phenomenon related to the blurring effect.

We also present a curve in Figure 6.12 that shows the Total Error Rate (TER) plotted as a function of the τ and $\varnothing$ parameters for the MOD method. Similar to PETS 2007 and PETS 2009, the optimal $\varnothing$ resemble those found in other works of the literature [31]. The curve follows the same trend as that of the earlier two datasets. The results obtained are illustrated with explanations in Figure 6.14 for MOD and Figure 6.13 for OSM.

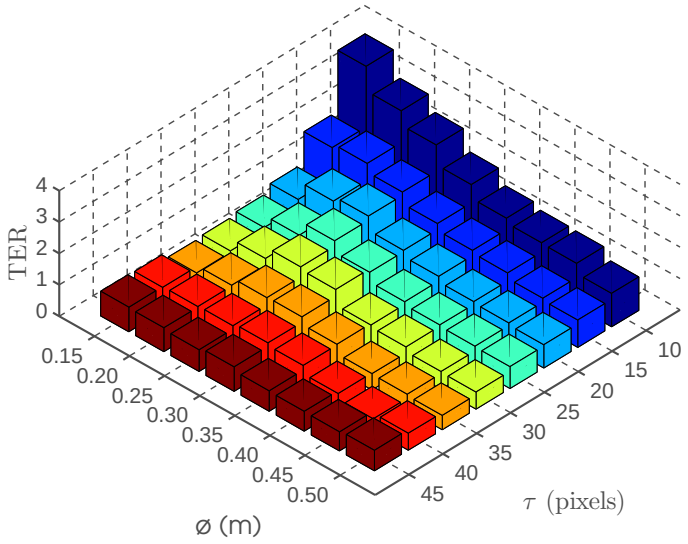


Figure 6.12: Evaluation of the MOD method with different parameter settings. Total Error Rate (TER) as a function of the τ and $\varnothing$ parameters for PETS 2007.

6.5 Summary

We can now summarize the results of all the experiments. For this purpose, we choose the Total Error Rate (TER) measure. We select it because this measure

Table 6.3: Summary of the results obtained.

Evaluation Measure	**Method/ Dataset**	**MSPL [13]**	**POM [23]**	**3DMPP [58]**	**OSM**	**MOD**
Total Error Rate	PETS 2006	0.28	-	-	0.12	**0.10**
	PETS 2007	1.08	-	-	**0.17**	0.36
	PETS 2009	0.28	0.25	0.12	**0.08**	0.13

is the sum of the errors resulting from false detection, missed detection, and of multiple detections. For this purpose, we have three datasets (PETS 2006, PETS 2007, and PETS 2009), and five algorithms:

- the **Multi-camera Occupancy map Deconvolution (MOD)** approach introduced in Chapter 4.
- the **Occupancy Shape Model (OSM)** approach introduced in Chapter 5
- the **Multiple Scene Planes Localization (MSPL)** method [13] that is the most popular in literature.
- the Probabilistic Occupancy Map (POM) [23]
- the 3D Marked Point Process (3DMPP) model [58, 59]

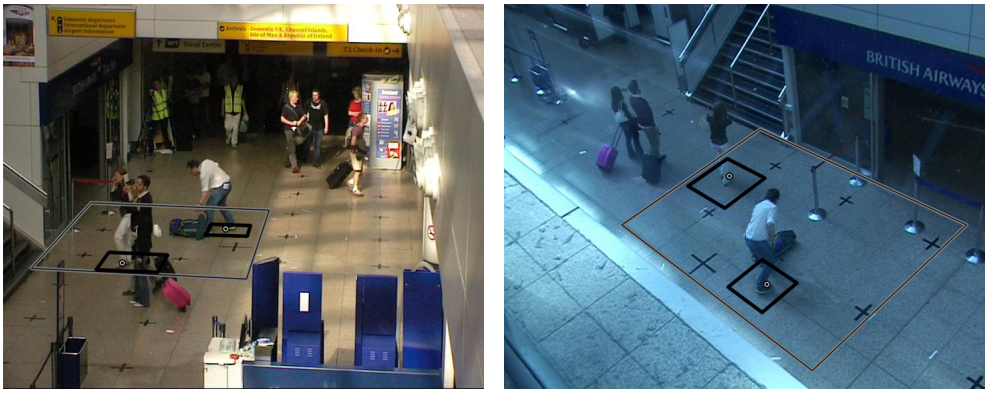
As we have obtained the POM and 3DMPP results from [58, 59], therefore, we are unable to show the performances of aforementioned algorithms on PETS 2006 and PETS 2007 datasets.

We can conclude from the Table 6.3 that MOD method surpasses MSPL, POM. MOD also provides comparable performances to 3DMPP. However, OSM remains the best method on the three databases. Moreover, OSM is also algorithmically efficient in terms of speed, and the number of parameters - in a sense, thanks to unsupervised learning, we only need to define and test one parameter. The reason for this success of OSM lies in multi-view 3D geometric reasoning coupled with the shape matching.

MOD tries to achieve a similar feat, shows success at 3D reasoning, but fails to handle the local deconvolutions by introducing a blur effect. Compared to MSPL, both methods perform multi-view reasoning developed around the ideas of ghost pruning. Finally, for extensive study of both proposed approaches, we made tests on three real, challenging, public datasets with varying vision-related complexities (shadows, lighting variations, density, clutter, noise), scenarios (indoor,

(a) PETS 2006 View 1

(b) PETS 2006 View 2



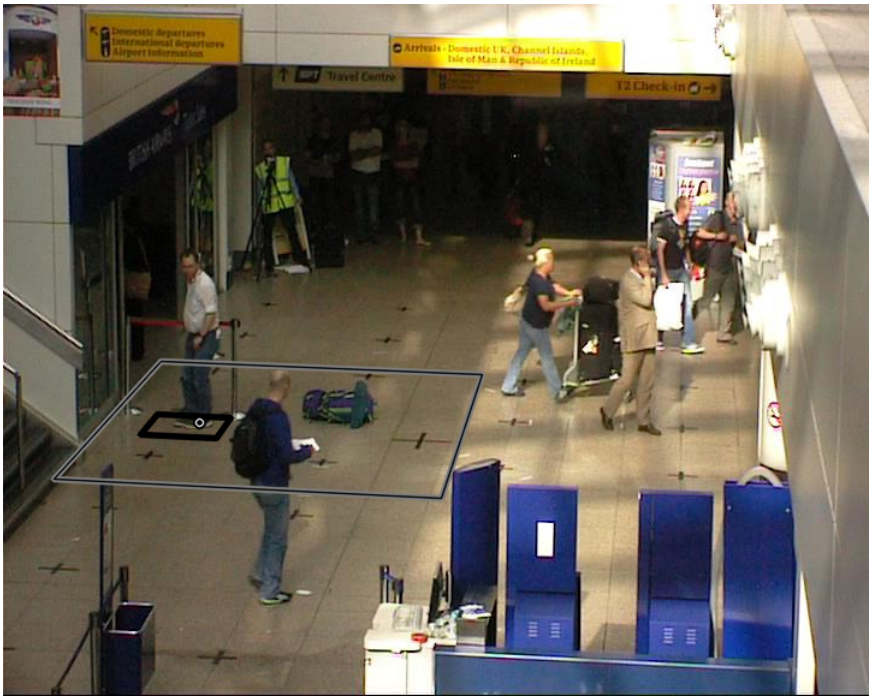
(c) PETS 2007 View 1

(d) PETS 2007 View 2

Figure 6.13: Examples of the estimated pedestrian locations in the PETS 2006 and PETS 2007 datasets (OSM). Notice the perspective effect in both the datasets. It is difficult to decide if the person is inside AOI or not. Nevertheless, the algorithm provides accurate detection.

outdoor, semi-indoor, semi-outdoor), camera equipment (color variations, compression levels), and configurations (camera placement at various levels of heights, uncontrolled).

We now conclude this thesis with a summary of our contributions. We then present the limitations of our work, and how we plan to improve on it in the future. This is discussed in the next chapter.



(a) PETS 2007
View 1

(b) PETS 2007
View 2

Figure 6.14: Example of the estimated pedestrian locations in the PETS 2007 datasets. The MOD algorithm correctly distinguishes between the pedestrians and bag, ambiguities of presence in Area of Interest (AOI) across views. The algorithm is also able to handle a mixture of indoor, outdoor situations, variations of intensity such as sunlight vs interior lighting, and the projections of pedestrians outside the user-defined AOI.

Chapter 7

Conclusions and Future Work

In this thesis, we have described a complete framework to perform people detection in images and videos obtained from multiple visual sensors. We have used the concepts of multi-view camera geometry, computer vision, and pattern recognition to propose two different approaches. We have presented the multi-camera occupancy maps and the ghost pruning phenomenon. The first approach is based on the deconvolution of the occupancy maps with a spatially varying kernel. In the second approach, we have proposed a keypoint extraction process in the occupancy maps. These keypoints are then validated by checking the presence of a particular shape model. We have tested these methods on three challenging datasets. We are able to perform robust people detection using multi-view reasoning in the multi-planar occupancy maps.

7.1 Summary of Key Contributions

- **Study of the multi-camera occupancy map and the process leading to ghost.** We have presented a detailed study of the multi-camera occupancy maps for performing people detection. Multi-camera occupancy is a popular technique for the fusion of information from multiple views projected into a common coordinate system. This technique has been applied in the literature to define probabilities of objects perpendicular to a plane, such as people standing in an airport or in a park. The occupancy maps perform this detection while exhibiting robustness to the conditions of varying illumination, shadows, resolution of the camera. However, this fusion technique

suffers from the phenomenon of ghosts. Ghosts are the false detections which occur by the overlap of several shapes in occupancy maps. These shapes are generated by the projections of the actual people. Ghosts create a false see-through effect and the erroneous detections. We have proposed a study for the multi-view reasoning methods to solve this without requiring temporal information. At the same time, we have presented robustness to color variations.

- **Efficient Modelling of Multi-camera Occupancy Map by a Convolution with a Spatially Varying Kernel.** We have proposed to study the process involved in the creation of the occupancy map. This led us to a mathematical modelling of the occupancy map by a convolution with spatially varying kernels that depend upon: (a) the 3D geometric shape present in the scene, (b) multi-view geometry (camera calibration).

- **People detection based on deconvolution.** We have proposed to approximate deconvolution by a similarity measure and obtain real positions of people in the scene by avoiding the ghost phenomena.

- **People detection based on a principle of hypothesis validation.** The previous approach suffers from a blurring effect introduced during the deconvolution process. In order to avoid this drawback, we first localize potential candidates using a keypoint extraction technique based on the watershed algorithm. Then, we accept or reject these hypothesis by checking the shape around them: a shape similar to that which would have been generated by a person at this position (from synthesis point of view) must be present. Further, we proposed an automated selection of the decision threshold leading to people detection.

- **Validation by experiments.** We have validated the proposed methods on three challenging datasets. We have also compared them with the other state-of-the-art algorithms to demonstrated the effectiveness of our approach.

7.2 Limitations of the Work

This section presents the limitations of our proposed work. There are several shortcomings of the proposed work: the following is a list of the intrinsic imperfections.

- **Blurred Effect of Deconvolution.** The deconvolution of the occupancy map by a spatially varying kernel produces a blur effect which makes the search of maxima difficult. This blur effect also causes the decrease in detection probabilities of the real objects. For the moment, we are unable to reduce the negative blurring effect for coarse-grained application of the convolutional framework. Therefore, we proposed another technique which could be understood as the limited scale application of the convolutional framework.

- **Keypoint Detection with Watershed Transform.** Multi-camera occupancy maps lead to significant values of interest corresponding to the locations of the object to be detected. There is also a particular shape present around these values. We have proposed a robust detection of the candidate detection locations or the keypoints based on the watershed transform. This particular keypoint detector has decreased the false detections and significantly reduced the number of multiple detections. However, this keypoint detector is designed around the idea of topological prominence. This topological prominence may not be conspicuously defined at the border regions. The kernel and its resulting convolutional framework can be designed to take into account the less topologically prominent pattern of values at the border e.g. by normalization. Alternatively, the tolerance threshold can be relaxed at the border areas. Furthermore, in our problem statement, we have to take into account the removal of projections originating from outside the AOI. Therefore, any improvement here will result in further reduction of Missed Detection Rate (MDR). As we have already shown significant decrease in the False Detection Rate (FDR), therefore, this will improve the overall performance of the multi-view detectors. Temporal information can potentially be useful in this particular case.

- **Height of the Visual Sensors.** Detection using multi-camera occupancy maps and the homography occupancy constraint are heavily influenced by

the camera heights. Homographic occupancy constraint defines the relation between the uncertainties of the image and the real world. A hypothetical camera or a set of cameras providing the top-down view of the scene will have a higher confidence in localizing the object but at the same time face uncertainty in the defining the height of the object. From the other extreme: if the same set of hypothetical cameras are placed lower to the ground then the definition of height becomes easier, but, the localization becomes an exercise of imprecise estimation. This effect of height on occupancy maps has been presented with experimental analysis and quantification in [31]. Features such as the appearance-based, texture-based, as in person re-identification, could be beneficial in this case [37].

In addition to the intrinsic deficiencies, there are also extensions and open issues related to our thesis which are presented in the future work section below.

7.3 Future Work

This section presents ideas which could result in the extension of our propositions. The list also includes improvements which could address the open issues faced in the field of multi-sensor visual surveillance.

- **Further Exploration of 3D Shapes.** We present two ideas related to the 3D shape modeling aspects of our work.
  - **Cuboids for Vehicle Detection.** Further explorations can be performed in the selection of 3D primitives in relation to the objects of

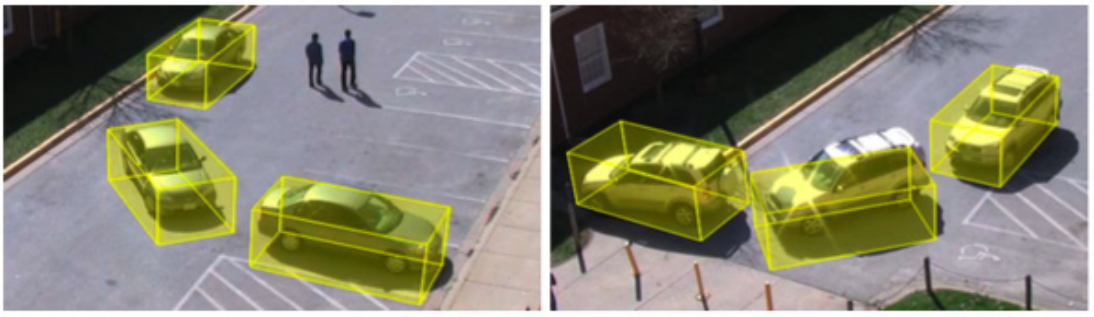


Figure 7.1: Vehicle Detection with 3D Cuboids. Reproduced from [31].

interest. For example, we can model vehicles using 3D cuboids. Vehicle detection is in active research at several laboratories including IFSTTAR. However, as cuboids are not radially symmetric like cylinders, all or specific orientations must be appropriately modelled. Bicycles may also be modeled with similar 3D primitives. This has been discussed in papers such as [31, 91].

  - **3D Modelling of the Occupancy Maps.** Occupancy maps are the fused projections of the camera views across multiple planes. It is possible to create a 3D plot of these multi-planar projections. One proposition could be to study this 3D plot and perform detection in it. This would change the problem statement from 2D shape analysis to 3D shape analysis. There is literature available on 3D human shape analysis which could be applied to such 3D occupancy maps [92–95].

- **Cross-sectional and Top-view.** Occupancy maps are fused projections that provide a 2D top-down view of the 3D real world. It could also be possible to generate occupancy maps from the four cross-sectional side views. Thus, occupancy maps could also be studied from the cross-sectional views. Furthermore, a fusion technique across both the cross-sectional and top-views may also be found. This will allow us to process the data using 2D techniques, and possibly further improve the detection rates.

- **Learning the Shapes in Occupancy Maps.** Shapes present in the occupancy maps can also be learned. For example, an *a priori* can be defined for the static objects or background in the scene such as the trees, lighting posts, signage, etc. This *a priori* could also help in adjusting the shape models, accounting for the particular scene. Similarly, the occupancy modeling can be further enhanced by modeling the presence of various people together. Currently, we are modeling the presence of one person in the scene at a time. This can be increased: multiple star-shaped structures or polygons can be synthesized. The particular distance across the shapes, their presence at a particular time at a particular spot (or not) could be learned and later applied for detection.

- **Temporal Domain.** Even though we have worked on detection, and proceed without using any temporal information; a simple intuitive extension is to apply the proposed work using consecutive frames, tracking, and to perform human recognition, behavior analysis [96].

- **Runtime.** We have not focused on hardcore runtime performances during this thesis. For MOD method, the spatially varying kernel and formation model exhibit linear dependence to the number of camera views and image resolution. The template similarity module has a linear dependency on the image resolution. For OSM method, the model generation and matching is significantly faster due to the fine-grained modeling and application only at the keypoints. For both OSM and MOD methods, the maxima or keypoint selection stage has a constant runtime. Some of the proposed algorithms have been implemented and tested using multi-core implementations. There is a significant potential for further computational improvements such as by utilizing only C++ codes, optimization for memory usage, and implementation of all parts with multi-core or GPU implementations.

Annexe A

Résumé en francais

Introduction

La détection de personnes dans les vidéos est un défi bien connu du domaine de la vision par ordinateur avec un grand nombre d'applications telles que le développement de systèmes de surveillance visuels. Même si les détecteurs monoculaires sont plus simples à mettre en place, ils ne sont pas adaptés dans le cadre de scènes complexes avec des occultations, une grande densité de personnes ou des scènes avec beaucoup de profondeur de champ menant à une grande variabilité dans la taille des personnes. Dans cette thèse, nous étudions la détection de personnes multi-caméra et plus particulièrement, l'utilisation de cartes d'occupation probabilistes créées en fusionnant les différentes vues grâce à la connaissance de la géométrie du système. La détection à partir de ces cartes d'occupation amène cependant à des fausses détections dues aux différentes projections. Celles-ci, bien connues dans la littérature, sont dénommées "fantôme". Aussi, nous proposons deux nouvelles techniques remédiant à ce problème et améliorant la détection des personnes. La première utilise une déconvolution par un noyau dont la forme varie spatialement tandis que la seconde est basée sur un principe de validation d'hypothèses. Ces deux approches n'utilisent volontairement pas l'information temporelle qui pourra être ré-introduite par la suite dans des algorithmes de suivi. Les deux approches ont été validées dans des conditions difficiles présentant des occultations, des encombrements plus ou moins denses et de fortes variations dans les réponses colorimétriques des caméras. Une comparaison avec d'autres méthodes

de l'état de l'art a également été menée sur trois bases de données publiques, validant les méthodes proposées dans le cadre des transports en commun, à savoir, la surveillance d'une gare et d'un aéroport.

Objectifs

Ce travail a pour objectif de détecter des personnes dans un contexte de vidéo-surveillance multi-caméras. Ces caméras sont fixes, observent la même scène sous différents points de vue, et sont reliées par des homographies supposées connues (ou que l'on déterminera grâce au calibrage des caméras). De tels systèmes, qui requièrent aussi la synchronisation des caméras, sont courants dans le domaine de la vidéo-surveillance. Néanmoins, la détection et le suivi de plusieurs personnes, dans un environnement non contraint reste un problème encore non encore résolu aujourd'hui et nous proposons, dans cette thèse, de repousser les limites des systèmes de détection de personnes multi-caméras. Ce problème est souvent considéré comme un problème de fusion de données dans un contexte multi-capteurs et nous proposons ici de le reformuler sous la forme d'un problème de reconnaissance des formes. En effet, ceci devrait permettre d'obtenir des résultats plus robustes dans des conditions particulièrement difficiles comme des caméras de différentes résolutions, de forts changements de point de vue entre caméras, des réponses colorimétriques différentes entre les caméras, des variations d'éclairage, du bruit, des occultations, des scènes plus ou moins denses etc. De plus, cette nouvelle formulation devrait également permettre d'éviter la détection de "fantômes", phénomène bien connu de la littérature en géométrie multi-vues.

Applications

L'être humain est un élément essentiel dans les interactions avec les machines et les interactions homme/machine ont de plus en plus d'importance dans la vie de tous les jours. Ainsi, la détection des personnes grâce à des capteurs visuels non intrusifs et l'utilisation de leur position amènent au développement de nombreuses applications qui jouent un rôle important dans la société. Ces applications sont en constante évolution. Nous en présentons ci-dessous une liste non exhaustive limitée au domaine de la vidéosurveillance.

Domaine des transports : Plusieurs applications utilisant la détection de personnes relèvent du domaine des transports comme la sécurité dans les trains, les bus, les gares ou les aéroports.

Domaine de la vie personnelle : La détection de personne est aussi utilisée dans le cadre des habitations personnelles avec par exemple des systèmes de surveillance à distance de domicile ou des systèmes de vidéosurveillance de personnes âgées ou atteintes de la maladie d'Alzheimer. Des applications plus ludiques telles que les jeux vidéo ont aussi connu un essor important ces dernières années avec notamment l'arrivée de capteurs tels que la kinect.

Domaine commercial : La sécurité est aussi importante dans le commerce ou de nombreux actes de vol, de vandalisme ou d'agression ont régulièrement lieu. Des compagnies un peu plus High-tech proposent également de modéliser le comportement des usagers afin d'améliorer leur offre. Ceci peut être utilisé pour mieux agencer un magasin ou pour fluidifier les flux de personnes.

Domaine public : La sécurité des personnes est un problème important que tous les gouvernements s'attèlent à résoudre. Pour cela, les systèmes de vidéosurveillance à base de caméras sont un outil indispensable, comme le prouvent les nombreuses caméras installées dans les zones publiques telles que la rue, les parcs, les hôpitaux etc. Même si ces caméras sont surtout utilisées comme outil d'enregistrement pour le moment, le besoin de traitement automatique est crucial et la première étape de la chaine de vidéosurveillance consiste justement à détecter les personnes. Des applications existent également dans le domaine militaire avec le développement de robots autonomes, de drones ou de satellites de surveillance.

Domaine de la recherche : Les algorithmes de détection de personnes sont utilisés dans de nombreuses applications encore du domaine de la recherche comme l'identification ou la ré-identification, la reconnaissance d'activité, l'analyse du comportement, etc. Autant d'applications qui gagneront en robustesse avec l'amélioration de la détection de personne.

TABLE A.1: Résultats de détection sur PETS2009.

Séquence	**Méthode**	**TER**	**FDR**	**MDR**	**MIR**
PETS 2009	Without pruning	0.362	0.291	0.047	0.024

Résumé des travaux

Dans un premier temps, nous réintroduisons le concept de carte d'occupation. Ces cartes sont obtenues en projetant les vues des différentes caméras sur plusieurs plans parallèles au sol, de hauteurs différentes et en agrégeant les projections dans un seul plan. Ces cartes d'occupation amènent une grande robustesse dans des situations difficiles comme les occultations, les changements d'éclairage ou les ombres [13, 60]. Dans le contexte de la vidéosurveillance et de la détection de personne, cette technique est très populaire mais est réputée pour produire de nombreuses fausses détections nommées "fantômes" dans la littérature [14, 24]. Ces "fantômes" sont dus à une configuration particulière de la position des objets et des caméras qui fait que les projections des différentes caméras s'intersectent à des positions particulières qui ne correspondent pas à une vraie personne. Table A.1, nous présentons les résultats de détection obtenus en recherchant les pics de la carte d'occupation sur une séquence de la littérature : la séquence PETS2009 [28]. FDR (false Detection Rate) correspond au taux de fausses détections, MDR (Missed Detection Rate) au taux de détection manquée, MIR (Multiple Instances Rate) au taux de multiples détections et TER (Total Erreur Rate) est l'erreur totale, définie comme la somme des 3 précédents taux. Ces taux ont été obtenus en choisissant le seuil de détection qui minimise l'erreur globale (TER).

Comme attendu, de nombreuses fausses détections sont obtenues (FDR est important). Durant cette thèse, nous proposons deux méthodes permettant de s'affranchir du problème des "fantômes".

Première approche proposée fondée sur une déconvolution

Dans un premier temps, nous avons étudié en détail le processus de création de ces "fantômes" et plus particulièrement, le processus de génération de la carte d'occupation. Ceci nous a permis de proposer une modélisation mathématique de la carte d'occupation sous la forme d'une convolution d'impulsions de Dirac (situées à la position des objets) et de noyaux dont la forme varie spatialement. En effet, leur forme dépend à la fois des propriétés de l'objet (forme et position) mais aussi, de la géométrie du système. Ceci permet de formaliser le problème de

détection comme une déconvolution dont le but sera de retrouver l'ensemble des impulsions de Dirac. Comme une vraie déconvolution n'est pas envisageable (la forme du noyau variant de pixel en pixel, il faudrait faire autant de déconvolution qu'il y a de pixel dans la carte d'occupation), nous proposons une approximation de la déconvolution en utilisant des mesures de similarités. Cette approche a été validée sur la base de données PETS2009 et comparée à cinq autres méthodes de la littérature qui suppriment les "fantômes" :

- Deux méthodes proposées par Ren et al (Ren eq 6 and Ren eq 7). [14]
- Un travail utilisant la carte d'occupation (Probabilistic Occupancy Map) POM [23]
- Un processus utilisant des points 3D (3D Marked Point Process) 3DMPP [58, 59]
- Une méthode utilisant plusieurs plans (Multiple Scene Plane Localization Method) MSPL [13]

Des résultats très encourageants ont été obtenus comme montré table A.2. Cette nouvelle méthode permet notament de diminuer considérablement le taux de fausses détection.

La méthode est cependant couteuse en temps de calcul et introduit un flou lors de la déconvolution qui rend ensuite la détection des pics difficile. Aussi, et en conservant le formalisme lié à la création de la carte d'occupation, une seconde méthode a été introduite.

Seconde approche proposée pour la suppression des fantômes de la carte d'occupation

TABLE A.2: Résultat de la première méthode utilisant une déconvolution.

Séquence	**Méthode**	**TER**	**FDR**	**MDR**	**MIR**
PETS 2009	Ren eq6 [14]	0.88	0.20	0.68	0.00
	Ren eq7 [14]	0.89	0.20	0.69	0.00
	MSPL [13]	0.28	0.18	0.10	0.00
	POM [23]	0.25	0.18	0.07	0.000
	3DMPP [58]	0.12	0.02	0.10	0.00
	Proposed	0.13	0.03	0.10	0.00

TABLE A.3: Résultats de la seconde méthode

Séquence	**Méthode**	**TER**	**FDR**	**MDR**	**MIR**
PETS 2009	MSPL [13]	0.28	0.18	0.10	0.00
	POM [23]	0.25	0.18	0.07	0.00
	3DMPP [58]	0.12	0.02	0.10	0.00
	Proposed	0.08	0.03	0.05	0.00

Cette seconde approche utilise la même modélisation que la première où une personne produit des accumulations dans la carte d'occupation mais aussi une forme spécifique autour de sa position (nommée noyaux précédemment). Ainsi, pour détecter une personne de manière robuste, nous proposons de sélectionner des candidats potentiels en recherchant les maximas locaux de la carte d'occupation puis de vérifier, pour chacun des candidats que le noyau, spécifique à la position de détection, est bien présent.

Cette méthode a été comparée à trois autres méthodes de la littérature :

- POM [23]
- 3DMPP [58]
- MSPL [13]

Les résultats, présentés Tableau A.3 montrent que non seulement la méthode est plus rapide que la précédente, mais en plus, elle en améliore les résultats grâce à une détection de maximas locaux plus facile (le flou introduit par la déconvolution a été évité)

Validation expérimentale

Les deux algorithmes proposés ont ensuite été validés sur deux autres bases de données de la littérature : PETS 2006 [26] and PETS 2007 [27]. Des résultats similaires à ceux obtenus sur PETS 2009 ont été trouvés et sur ces trois séquences, les approches proposées dépassent l'état de l'art.

Conclusion et perspectives

Le document se termine par une conclusion générale, un bilan des limites de l'approche proposée et donne les perspectives qui pourraient être envisagée pour poursuivre ce travail.

Bibliography

[1] LakhmiC. Jain and CheePeng Lim. Advances in intelligent methodologies and techniques. In Horia-Nicolai Teodorescu, Junzo Watada, and LakhmiC. Jain, editors, *Intelligent Systems and Technologies*, volume 217 of *Studies in Computational Intelligence*, pages 3–28. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-01884-8. doi: 10.1007/978-3-642-01885-5_1. URL http://dx.doi.org/10.1007/978-3-642-01885-5_1.

[2] Marcus Nieto. Public video surveillance: Is it an effective crime prevention tool? http://www.library.ca.gov/CRB/97/05/, 1997. [Online].

[3] CISCO. Virtual patient observation: Centralize monitoring of high-risk patients with video. http://www.cisco.com/c/en/us/products/collateral/physical-security/video-surveillance-manager/white_paper_C11-715263.html, 2015. [Online].

[4] VideoSurveillance.com LLC. Video surveillance for child care. http://www.videosurveillance.com/child-care.asp, 2015. [Online].

[5] Hamid Aghajan and Andrea Cavallaro. *Multi-Camera Networks: Principles and Applications*. Academic Press, 2009. ISBN 0123746337, 9780123746337.

[6] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003. ISBN 0521540518.

[7] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recogn. Lett.*, 34(1):3–19, January 2013. ISSN 0167-8655. doi: 10.1016/j.patrec.2012.07.005. URL http://dx.doi.org/10.1016/j.patrec.2012.07.005.

[8] Martijn Liem and Dariu Gavrila. International conference on computer vision systems 2013 dataset. http://www.gavrila.net/Datasets/Univ_

_of_Amsterdam_Multi-Cam_P/UvA_Multi-Camera_Multi-Person_/uva_multi-camera_multi-person_.html, 2013. [Online].

[9] Francois Bremond, Anh Tuan Nghiem, Valery Valentin, Ruihua Ma, and Monique Thonnat. Etiseo - an evaluation programme for video surveillance algorithms. http://www-sop.inria.fr/members/Francois.Bremond/topicsText/etiseoProject.html, 2007. [Online].

[10] H. B. Mitchell. *Multi-Sensor Data Fusion: An Introduction.* Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 3540714634, 9783540714637.

[11] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376, December 1992. ISSN 0360-0300. doi: 10.1145/146370.146374. URL http://doi.acm.org/10.1145/146370.146374.

[12] Saad Masood Khan and Mubarak Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *In European Conference on Computer Vision*, 2006.

[13] Saad M. Khan and Mubarak Shah. Tracking multiple occluding people by localizing on multiple scene planes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):505–519, March 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.102.

[14] Jie Ren, Ming Xu, and J.S. Smith. Pruning phantom detections from multiview foreground intersection. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1025–1028, Sept 2012. doi: 10.1109/ICIP.2012.6467037.

[15] Ran Eshel and Yael Moses. Tracking in a dense crowd using multiple cameras. *Int. J. Comput. Vision*, 88(1):129–143, May 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0307-0. URL http://dx.doi.org/10.1007/s11263-009-0307-0.

[16] CCTV Camera Pros. Laundromat surveillance system. remote dvr viewer. http://www.cctvcamerapros.com/Laundromat-Surveillance-System-s/272.htm, 2015. [Online].

[17] CISCO. Transportation agency centralizes video surveillance. http://www.cisco.com/c/en/us/products/collateral/physical-security/

video-surveillance-manager/casestudy_c36-610593.html, 2015. [Online].

[18] Microsoft. Kinect for xbox one. http://www.xbox.com/en-US/xbox-one/accessories/kinect-for-xbox-one, 2015. [Online].

[19] Microsoft. Kinect for windows. https://www.microsoft.com/en-us/kinectforwindows/meetkinect/default.aspx, 2015. [Online].

[20] Google Research. Building a deeper understanding of images. http://googleresearch.blogspot.fr/2014/09/building-deeper-understanding-of-images.html, 2015. [Online].

[21] National Institute of Justice. Detection and surveillance technologies. http://www.nij.gov/topics/technology/detection-surveillance/pages/welcome.aspx, 2015. [Online].

[22] Bernt Schiele, Mykhaylo Andriluka, Nikodem Majer, Stefan Roth, and Christian Wojek. Visual people detection: Different models, comparison and discussion. In *Proceedings of the IEEE ICRA Workshop on People Detection and Tracking*, May 2009.

[23] François Fleuret, Jérôme Berclaz, Richard. Lengagne, and Patrick Fua. Multicamera people tracking with a probabilistic occupancy map. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):267–282, Feb 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1174.

[24] Murray Evans, Longzhen Li, and James Ferryman. Suppression of detection ghosts in homography based pedestrian detection. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE*, pages 31–36, Sept 2012. doi: 10.1109/AVSS.2012.73.

[25] Zoltan Megyesi. *Dense Matching Methods for 3D Scene Reconstruction from Wide Baseline Images*. PhD thesis, Eotvos Lorand University, 2009.

[26] PETS. Performance evaluation of tracking and surveillance dataset 2006. http://www.cvg.reading.ac.uk/PETS2006/data.html, 2006. [Online].

[27] PETS. Performance evaluation of tracking and surveillance dataset 2007. http://www.cvg.reading.ac.uk/PETS2007/data.html, 2007. [Online].

[28] PETS. Pets dataset: Performance evaluation of tracking and surveillance. http://www.cvg.rdg.ac.uk/PETS2009/a.html, 2009. [Online].

[29] Seungmin Rho, Geyong Min, and Weifeng Chen. Advanced issues in artificial intelligence and pattern recognition for intelligent surveillance system in smart home environment. *Engineering Applications of Artificial Intelligence*, 25(7):1299 – 1300, 2012. ISSN 0952-1976. doi: http://dx.doi.org/10.1016/j.engappai.2012.07.007. URL http://www.sciencedirect.com/science/article/pii/S0952197612001959. Advanced issues in Artificial Intelligence and Pattern Recognition for Intelligent Surveillance System in Smart Home Environment.

[30] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, 2012. URL http://dblp.uni-trier.de/db/journals/pami/pami34.html#DollarWSP12.

[31] Peter Carr, Yaser Sheikh, and Iain Matthews. Monocular object detection using 3d geometric primitives. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, volume 7572 of *Lecture Notes in Computer Science*, pages 864–878. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-33717-8. doi: 10.1007/978-3-642-33718-5_62. URL http://dx.doi.org/10.1007/978-3-642-33718-5_62.

[32] Markus Enzweiler and Dariu M. Gavrila. Monocular pedestrian detection: Survey and experiments. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2179–2195, Dec 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.260.

[33] David Gerónimo, Antonio M. López, and Thorsten Graf. Survey on pedestrian detection for advanced driver assistance systems. In *IEEE PAMI, available online: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.122*, 2009.

[34] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4), December 2006. ISSN 0360-0300. doi: 10.1145/1177352.1177355. URL http://doi.acm.org/10.1145/1177352.1177355.

[35] Yali Amit and Pedro Felzenszwalb. Object detection. In *Computer Vision: A Reference Guide*.

[36] Amit K Roy Chowdhury. Distributed camera network. http://motion.me.ucsb.edu/SBControlWorkshop-24jun2011/PDFs/Roy-Chowdhury-SBCWorkshop2011.pdf, 2015. [Online].

[37] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Comput. Surv.*, 46(2):29:1–29:37, December 2013. ISSN 0360-0300. doi: 10.1145/2543581.2543596. URL http://doi.acm.org/10.1145/2543581.2543596.

[38] Jake K. Aggarwal and Michael S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, April 2011. ISSN 0360-0300. doi: 10.1145/1922649.1922653. URL http://doi.acm.org/10.1145/1922649.1922653.

[39] T. A. Clarke and J. G. Fryer. The development of camera calibration methods and models. *The Photogrammetric Record*, 16(91):51–66, 1998. ISSN 1477-9730. doi: 10.1111/0031-868X.00113. URL http://dx.doi.org/10.1111/0031-868X.00113.

[40] Paulo Dias. Tsai camera calibration. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/DIAS1/, 2015. [Online].

[41] Yiannis Aloimonos. Camera calibration. https://www.cs.umd.edu/class/fall2013/cmsc426/lectures/camera-calibration.pdf, 2015. [Online].

[42] Richard Szeliski. Camera calibration and pose estimation. http://courses.cs.washington.edu/courses/cse590ss/01wi/notes/CalBundleLecture5g.pdf, 2015. [Online].

[43] Roger Y. Tsai. Radiometry. chapter A Versatile Camera Calibration Technique for High-accuracy 3D Machine Vision Metrology Using Off-the-shelf TV Cameras and Lenses, pages 221–244. Jones and Bartlett Publishers, Inc., USA, 1992. ISBN 0-86720-294-7. URL http://dl.acm.org/citation.cfm?id=136913.136938.

[44] Zhengyou Zhang. Camera calibration. In *Emergin Topics in Computer Vision*, chapter 2, pages 4–43. Prentice Hall Professional Technical Reference, 2004.

[45] A. Basu. Active calibration of cameras: theory and implementation. *Systems, Man and Cybernetics, IEEE Transactions on*, 25(2):256–265, Feb 1995. ISSN 0018-9472. doi: 10.1109/21.364838.

[46] Xavier Armangué, Joaquim Salvi, and Joan Batlle. A comparative review of camera calibrating methods with accuracy evaluation. *Pattern Recognition*, 35:1617–1635, 2000.

[47] Liming Song, Wenfu Wu, Junrong Guo, and Xiuhua Li. Survey on camera calibration technique. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013 5th International Conference on*, volume 2, pages 389–392, Aug 2013. doi: 10.1109/IHMSC.2013.240.

[48] Yaser Ajmal Sheikh and Mubarak Shah. Trajectory association across multiple airborne cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):361–367, February 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.70750. URL http://dx.doi.org/10.1109/TPAMI.2007.70750.

[49] Anthony Whitehead, Robert Laganiere, and Prosenjit Bose. Temporal synchronization of video sequences in theory and in practice. In *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, volume 2, pages 132–137, Jan 2005. doi: 10.1109/ACVMOT.2005.114.

[50] Jeremy Eric Elson and Deborah L. Estrin. Time synchronization in wireless sensor networks. In *International Parallel and Distributed Processing Symposium (IPDPS), Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing*, april 2001.

[51] Daniel G. Costa, Ivanovitch Silva, Luiz Affonso Guedes, Francisco Vasques, and Paulo Portugal. Availability issues in wireless visual sensor networks. *Sensors*, 14(2):2795, 2014. ISSN 1424-8220. doi: 10.3390/s140202795. URL http://www.mdpi.com/1424-8220/14/2/2795.

[52] Xiaotao Zou, B. Bhanu, Bi Song, and A.K. Roy-Chowdhury. Determining topology in a distributed camera network. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 5, pages V – 133–V – 136, Sept 2007. doi: 10.1109/ICIP.2007.4379783.

[53] Richard J. Radke. Multiview geometry for camera networks. In *Multi-Camera Networks: Concepts and Applications*. Elsevier, 2009.

[54] F. Schaffalitzky. *Grouping, Matching and Reconstruction in Multiple View Geometry*. PhD thesis, University of Oxford, apr 2002.

[55] N. Anjum, M. Taj, and A. Cavallaro. Relative position estimation of non-overlapping cameras. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–281–II–284, April 2007. doi: 10.1109/ICASSP.2007.366227.

[56] Kinh Tieu, G. Dalley, and W.E.L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1842–1849 Vol. 2, Oct 2005. doi: 10.1109/ICCV.2005.122.

[57] M. Taj and A. Cavallaro. Distributed and decentralized multicamera tracking. *Signal Processing Magazine, IEEE*, 28(3):46–58, May 2011. ISSN 1053-5888. doi: 10.1109/MSP.2011.940281.

[58] Akos Utasi and Csaba. Benedek. A bayesian approach on people localization in multicamera systems. *Circuits and Systems for Video Technology, IEEE Transactions on*, 23(1):105–115, Jan 2013. ISSN 1051-8215. doi: 10.1109/TCSVT.2012.2203201.

[59] A Utasi and C. Benedek. A 3-d marked point process model for multi-view people detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3385–3392, June 2011. doi: 10.1109/CVPR.2011.5995699.

[60] Ran Eshel and Yael Moses. Tracking in a dense crowd using multiple cameras. *International Journal of Computer Vision*, 88(1):129–143, 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0307-0. URL http://dx.doi.org/10.1007/s11263-009-0307-0.

[61] C. Benedek and Tamas Sziranyi. Bayesian foreground and shadow detection in uncertain frame rate surveillance videos. *Image Processing, IEEE Transactions on*, 17(4):608–621, April 2008. ISSN 1057-7149. doi: 10.1109/TIP.2008.916989.

[62] Weina Ge and Robert T. Collins. Crowd detection with a multiview sampler. In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pages 324–337, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15554-5, 978-3-642-15554-3. URL http://dl.acm.org/citation.cfm?id=1888150.1888177.

[63] Katerina Fragkiadaki and Jianbo Shi. Figure-ground image segmentation helps weakly-supervised learning of objects. In *Proceedings of the 11th European conference on Computer vision: Part VI*, ECCV'10, pages 561–574, 2010.

[64] Muhammad Owais Mehmood, Sebastien Ambellouis, and Catherine Achard. Ghost pruning for people localization in overlapping multicamera systems. In *VISAPP (2)*, pages 632–639, 2014.

[65] James Orwell, Douglas S. Massey, Paolo Remagnino, Stuart D. Greenhill, and Graeme A. Jones. A multi-agent framework for visual surveillance. In *Image Analysis and Processing, 1999. Proceedings. International Conference on*, pages 1104–1107, 1999. doi: 10.1109/ICIAP.1999.797748.

[66] John Krumm, Steve Harris, Brian Meyers, Barry Brumitt, Michael Hale, and Steve Shafer. Multi-camera multi-person tracking for easyliving. In *Visual Surveillance, 2000. Proceedings. Third IEEE International Workshop on*, pages 3–10, 2000. doi: 10.1109/VS.2000.856852.

[67] Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, pages –252 Vol. 2, 1999. doi: 10.1109/CVPR.1999.784637.

[68] Chirstopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex P. Pentland. Pfinder: real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):780–785, Jul 1997. ISSN 0162-8828. doi: 10.1109/34.598236.

[69] Jian Yao and Jean-Marc Odobez. Multi-layer background subtraction based on color and texture. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.

[70] Hasan Soltanzadeh and Mohammad Rahmati. Recognition of persian handwritten digits using image profiles of multiple orientations. *Pattern Recogn. Lett.*, 25(14):1569–1576, October 2004. ISSN 0167-8655. doi: 10.1016/j.patrec.2004.05.014. URL http://dx.doi.org/10.1016/j.patrec.2004.05.014.

[71] Jian Yao and Jean-Marc Odobez. Multi-layer background subtraction based on color and texture. In *Computer Vision and Pattern Recognition, 2007.*

CVPR '07. IEEE Conference on, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383497.

[72] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/0031-3203(95)00067-4. URL http://www.sciencedirect.com/science/article/pii/0031320395000674.

[73] Marco Cristani, Michela Farenzena, Domenico Bloisi, and Vittorio Murino. Background subtraction for automated multisensor surveillance: A comprehensive review. *EURASIP Journal on Advances in Signal Processing*, 2010(1):343057, 2010. ISSN 1687-6180. doi: 10.1155/2010/343057. URL http://asp.eurasipjournals.com/content/2010/1/343057.

[74] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991. ISSN 0920-5691. doi: 10.1007/BF00130487. URL http://dx.doi.org/10.1007/BF00130487.

[75] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In Peter Auer and Ron Meir, editors, *Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 545–560. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-26556-6. doi: 10.1007/11503415_37. URL http://dx.doi.org/10.1007/11503415_37.

[76] MathWorks Steve Eddins. The watershed transform: Strategies for image segmentation. http://fr.mathworks.com/company/newsletters/articles/the-watershed-transform-strategies-for-image-segmentation.html, 2015. [Online].

[77] Serge Beucher and Fernand Meyer. The morphological approach to segmentation: the watershed transformation. Mathematical morphology in image processing. *Optical Engineering*, 34:433–481, 1993.

[78] Osman Topçu, Ali Özer Ercan, and A. Aydin Alatan. Recovery of temporal synchronization error through online 3d tracking with two cameras. In *Proceedings of the International Conference on Distributed Smart Cameras*, ICDSC '14, pages 27:1–27:6, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2925-5. doi: 10.1145/2659021.2659056. URL http://doi.acm.org/10.1145/2659021.2659056.

[79] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. ISSN 1931-9193. doi: 10.1002/nav.3800020109. URL http://dx.doi.org/10.1002/nav.3800020109.

[80] Muhammad Owais Mehmood, Sébastien Ambellouis, and Catherine Achard. Launch these Manhunts! Shaping the Synergy Maps for Multi-Camera Detection. In *VISAPP, International Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, page 8p, Berlin, March 2015.

[81] Jingchen Liu, Robert T. Collins, and Yanxi Liu. Robust autocalibration for a surveillance camera network. *IEEE Winter Conference on Applications of Computer Vision*, 0:433–440, 2013. ISSN 1550-5790. doi: http://doi.ieeecomputersociety.org/10.1109/WACV.2013.6475051.

[82] Anil K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, September 1999. ISSN 0360-0300. doi: 10.1145/331499.331504. URL http://doi.acm.org/10.1145/331499.331504.

[83] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

[84] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

[85] Yuan Li, Bo Wu, and R. Nevatia. Human detection by searching in 3d space using camera and scene knowledge. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–5, Dec 2008. doi: 10.1109/ICPR.2008.4761709.

[86] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1 edition, September 1992. ISBN 0471547700. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471547700.

[87] Chris Fraley and Adrian E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classif.*, 24(2):155–181,

September 2007. ISSN 0176-4268. doi: 10.1007/s00357-007-0004-5. URL http://dx.doi.org/10.1007/s00357-007-0004-5.

[88] Andrews Sobral. BGSLibrary: An opencv c++ background subtraction library. In *IX Workshop de Visão Computacional (WVC'2013)*, Rio de Janeiro, Brazil, Jun 2013. URL https://github.com/andrewssobral/bgslibrary.

[89] Ákos Kiss and Tamás Szirányi. Localizing people in multi-view environment using height map reconstruction in real-time. *Pattern Recognition Letters*, 34(16):2135 – 2143, 2013. ISSN 0167-8655. doi: http://dx.doi.org/10.1016/j.patrec.2013.08.007. URL http://www.sciencedirect.com/science/article/pii/S0167865513003024.

[90] Peter A Flach. Roc analysis. *Encyclopedia of machine learning*, pages 869–875, 2010.

[91] M. Zeeshan Zia, Michael Stark, Bernt Schiele, and Konrad Schindler. Detailed 3d representations for object recognition and modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2608–2623, 2013. doi: 10.1109/TPAMI.2013.87. URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.87.

[92] Rim Slama, Hazem Wannous, and Mohamed Daoudi. Extremal Human Curves: a New Human Body Shape and Pose Descriptor. In *10th IEEE International Conference on Automatic Face and Gesture Recognition*, page ., Shanghai, China, April 2013. URL https://hal.archives-ouvertes.fr/hal-00784488.

[93] Pengcheng Xi, Hongyu Guo, and Chang Shu. Human body shape prediction and analysis using predictive clustering tree. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, pages 196–203, May 2011. doi: 10.1109/3DIMPVT.2011.32.

[94] Greg Mori and Jitendra Malik. Recovering 3d human body configurations using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(7):1052–1062, July 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.149.

[95] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, June 2015.

[96] Salma Elloumi, Serham Cosar, Guido Pusiol, François. Bremond, and Monique. Thonnat. Unsupervised discovery of human activities from long-videos. *IET Computer Vision*, 2014.

Titre en français: Detection de personnes pour des systèmes de videosurveillance multi-caméra intelligents

Résumé en français: La détection de personnes dans les vidéos est un défi bien connu du domaine de la vision par ordinateur avec un grand nombre d'applications telles que le développement de systèmes de surveillance visuels. Même si les détecteurs monoculaires sont plus simples à mettre en place, ils sont dans l'incapacité de gérer des scènes complexes avec des occultations, une grande densité de personnes ou des scènes avec beaucoup de profondeur de champ menant à une grande variabilité dans la taille des personnes. Dans cette thèse, nous étudions la détection de personnes par un système multicaméras et plus particulièrement, l'utilisation de cartes d'occupation probabilistes créées en fusionnant les différentes vues grâce à la connaissance de la géométrie du système. La détection à partir de ces cartes d'occupation amène cependant de fausses détections dues aux différentes projections. Celles-ci, bien connues dans la littérature, sont dénommées "fantôme". Aussi, nous proposons deux nouvelles techniques remédiant à ce problème et améliorant la détection des personnes. La première utilise une déconvolution par un noyau dont la forme varie spatialement tandis que la seconde est basée sur un principe de validation d'hypothèse. Ces deux approches n'utilisent volontairement pas l'information temporelle qui pourra être ré-introduite par la suite dans des algorithmes de suivi. Les deux approches ont été validées dans des conditions difficiles présentant des occultations, des encombrements plus ou moins denses et de fortes variations dans les réponses colorimétriques des caméras. Une comparaison avec d'autres méthodes de l'état de l'art a également été menée sur trois bases de données publiques, validant les méthodes proposées dans le cadre des transports en commun, à savoir, la surveillance d'une gare et d'un aéroport.

Mots-clefs: Géométrie multi-vues, Fusion de capteurs, Reconnaissance des Formes, Détection d'objects, Surveillance.

Titre en anglais: People Detection Methods For Intelligent Multi-Camera Surveillance Systems

Résumé en anglais: People detection is a well-studied open challenge in the field of Computer Vision with applications such as in the visual surveillance systems. Monocular detectors have limited ability to handle occlusion, clutter, scale, density. Ubiquitous presence of cameras and computational resources fuel the development of multi-camera detection systems. In this thesis, we study the multi-camera people detection; specifically, the use of multi-view probabilistic occupancy maps based on the camera calibration. Occupancy maps allow multi-view geometric fusion of several camera views. Detection with such maps create several false detections and we study this phenomenon: ghost pruning. Further, we propose two novel techniques in order to improve multi-view detection based on: (a) kernel deconvolution, and (b) occupancy shape modeling. We perform non-temporal, multi-view reasoning in occupancy maps to recover accurate positions of people in challenging conditions such as of occlusion, clutter, lighting, and camera variations. We show improvements in people detections across three challenging datasets for visual surveillance including comparison with state-of-the-art techniques. We show the application of this work in exigent transportation scenarios i.e. people detection for surveillance at a train station and at an airport.

Mots-clefs: Multi-view Geometry, Sensor Fusion, Pattern Recognition, Object Detection, Surveillance.