THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Occlusion Handling in Video Surveillance Systems

Ali Emami

A thesis submitted for the degree of Doctor of Philosophy at The University of Queensland in 2015

School of Information Technology and Electrical Engineering

Abstract

The major focus of this thesis is on the problem of occlusion handling in monocular visual tracking and pedestrian detection for video surveillance applications. Despite all the progress in this field, robust and reliable tracking in real world scenarios, including interacting objects with frequent occlusion events, is still an open research problem. Various tracking systems require different strategies for resolving occlusion situations due to their different structural properties. The research described in this thesis advances state-of-the-art with two significant occlusion handling methods, corresponding with two typical visual tracking approaches: **(i)** Template based trackers; and **(ii)** Detection based trackers, relying on off-line trained detectors. Tracking performance in template-based trackers, is very dependent on a valid up-to-date target model. In the first part of the research, we propose an occlusion handling framework for template trackers, which provides a basis for protecting the target model against corruption in occlusion and drift situations. The occlusion model is built upon motion dynamics of the targets, described through multi-channel Spatio-temporal Oriented Energy features. The proposed model is used for identifying the occlusion mode of the target in the course of tracking, as well as estimating an occlusion mask to determine visible parts of the target in occlusion events. By focusing on visible parts of the targets for template matching, an improved tracking performance is obtained. In the second part, we introduce an efficient data association method for multiple pedestrian tracking by detection, through a light-weight spatiotemporal clustering framework. Such tracking approaches typically depend on off-line trained object detectors, while tracking is performed by data association among the detection results. Our proposed scheme resolves the occlusion and confusion ambiguities among the interacting targets and simultaneously compensates for the general detection errors, such as missed or false detections. The suggested method may be introduced as an improved adaptive Non-Max-Suppression (NMS) method, which is aware of the number of existing targets in the scene and provides a solid performance for highly overlapped targets. The framework demonstrates a real-time performance with high capability of occlusion handling in low resolution imagery.

Declaration by Author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the *Copyright Act 1968* unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Journal Articles

1. Ali Emami, Mehrtash T. Harandi, Farhad Dadgostar, Brian C. Lovell. 'Novelty detection in human tracking based on spatiotemporal oriented energies'. In *Pattern Recognition*, Vol. 48, No. 3, pp. 812-826, 2015.

Conference Papers

1. Ali Emami, Abbas Bigdeli, Adam Postula. 'High Throughput Variable Size Non-square Gabor Engine with Feature Pooling Based on GPU'. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 393-399, 2010.

2. Ali Emami, Fargad Dadgostar, Abbas Bigdeli, Brian C. Lovell. 'Role of Spatiotemporal Oriented Energy Features for Robust Visual Tracking in Video Surveillance'. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pp. 349-354, 2012.

Patents

1. Ali Emami, Mehrtash T. Harandi, Lisa Brown, Sharath Pankanti. 'Real-time Occlusion Handling in Pedestrian Detection and Tracking'.
   Non-provisional patent filed in USPTO: PCT international application number 14/719,875 filed on May 22, 2015.
   (Joint research study between the IBM Watson Lab. in New York, USA and The University of Queensland in Australia, which was partially funded through a Graduate School International Travel Award.)

Publications included in the thesis

1. Ali Emami, Abbas Bigdeli, Adam Postula. 'High Throughput Variable Size Non-square Gabor Engine with Feature Pooling Based on GPU'. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 393-399, 2010.
   Incorporated in Chapter 5.

Contributor	*Statement of contribution*
Ali Emami (Candidate)	Conception and design of algorithm (100%) Design of experiments (90%) Paper writing (85%)
Abbas Bigdeli	Design of experiments (10%) Paper writing and editing (10%)
Adam Postula	Paper editing and proof reading (5%)

2. Ali Emami, Fargad Dadgostar, Abbas Bigdeli, Brian C. Lovell. 'Role of Spatiotemporal Oriented Energy Features for Robust Visual Tracking in Video Surveillance'. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pp. 349-354, 2012.
   Incorporated in Chapter 3.

Contributor	*Statement of contribution*
Ali Emami (Candidate)	Conception and design of algorithm (90%) Design of experiments (90%) Paper writing (80%)
Farhad Dadgostar	Conception and design of algorithm (10%) Design of experiments (10%) Paper writing and editing (10%)
Abbas Bigdeli	Paper editing and proof reading (5%)
Brian C. Lovell	Paper editing and proof reading (5%)

3. Ali Emami, Mehrtash T. Harandi, Farhad Dadgostar, Brian C. Lovell. 'Novelty detection in human tracking based on spatiotemporal oriented energies'. In *Pattern Recognition*, Vol. 48, No. 3, pp. 812-826, 2015.
Incorporated in Chapter 3.

Contributor	*Statement of contribution*
Ali Emami (Candidate)	Conception and design of algorithm (90%) Design of experiments (90%) Paper writing (80%)
Mehrtash T. Harandi	Design of experiments (5%) Paper writing and editing (10%)
Farhad Dadgostar	Conception and design of algorithm (10%) Design of experiments (5%) Paper writing and editing (5%)
Brian C. Lovell	Paper editing and proof reading (5%)

4. Ali Emami, Mehrtash T. Harandi, Lisa Brown, Sharath Pankanti. 'Real-time Occlusion Handling in Pedestrian Detection and Tracking'. **Non-provisional patent filed in USPTO: PCT international application number 14/719,875 filed on May 22, 2015.** (Joint research study between the IBM Watson lab and The University of Queensland)
Incorporated in Chapter 4.

Contributor	*Statement of contribution*
Ali Emami (Candidate)	Conception and design of algorithm (90%) Design of experiments (90%) Disclosure report and presentation (80%)
Mehrtash T. Harandi	Conception and design of algorithm (5%) Design of experiments (5%) Disclosure report, writing and editing (10%)
Lisa Brown	Conception and design of algorithm (5%) Design of experiments (5%) Disclosure report, editing and proof reading (5%)
Sharath Pankanti	Disclosure report and presentation editing (5%)

Contributions by Others to the Thesis

The work contained in this thesis was carried out by the author under the guidance and supervision of his advisors, Prof. Brian C. Lovell, Dr Abbas Bigdeli and Dr Farhad Dadgostar. The research included in Chapter 4 of the thesis, which leaded to filing a non-provisional patent in USPTO, was accomplished by the author in IBM Watson Lab. under the supervision of Prof. Sharath Pankanti and Dr Lisa Brown, within a joint research study between the IBM Watson Lab. and The University of Queensland. Part of the work contained in this thesis was carried out by the author under the collaboration and discussions with Dr. Mehrtash Harandi.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgments

I want to especially thank my advisory team, Prof. Brian Lovell, Dr Abbas Bigdeli and Dr Farhad Dadgostar for their mentorship throughout my PhD research. They have been very supportive, understanding and patient throughout the entire process, providing professional guidance and kind encouragement to help shape my research career. I am honored to be a member of the 'Advanced Video Surveillance Group-AVSG' directed by Prof. Brian Lovell and to be a member of NICTA QLD Research Lab. I have picked a lot in our weekly group meetings, throughout the scientific discussions with our colleagues. I thank all my friends in AVSG and NICTA, to give me a great field work experience as a computer scientist.

I would also like to give my special appreciation to Dr Mehrtash Harandi, from whom I have learned a vast amount of knowledge not only in the areas of computer vision and analytical development, but also in technical writing and presentation techniques. Many thanks to Prof Sharath Pankanti and Dr Lisa Brown in 'Exploratory Computer Vision' group of IBM Watson Laboratory, who provided me a unique opportunity to participate as a research scholar in the cutting edge on-going research within the IBM community. This has been one of the most fruitful periods throughout my research life ever.

I acknowledge the financial support granted to me through the University of Queensland Scholarship and the top-up Scholarship of the NICTA research center, as well as the 'International Travel Awards' from UQ and NICTA that provided me the opportunity to participate in international conferences and internships. The joint research study between UQ and IBM, which was based in IBM Watson laboratory in New-York city and leaded to filing a US patent, was partially funded through a UQ Graduate School International Travel Award (GSITA).

To my family, my beloved mother and father and my dear brothers, I cannot thank you enough for your endless support and encouragement. Even though we are geographically separated, I know you have always been sending your best wishes and blessings to me. The knowledge of your support have always given me the courage to stand on my feet and progress in every

situation.

Finally to my dear wife, Laleh and my lovely daughter, Bahar. Thank you for your toleration throughout the years, even in my greatest moments of anxiety. Your understanding, encouragement, companionship and confidence in me always brought me strength to keep pushing forward. Living with a PhD student can not be easy, but you have always stood by me patiently and supportively throughout this journey. I am honored to have you by my side and look forward to a glorious future on our upcoming journeys together. Thanks to my dear daughter Bahar, who spent her childhood with various concerns and sacrificed some of her childhood dreams for the family and her dad.

Keywords

computer vision, machine learning, visual tracking, pedestrian detection, occlusion modeling, pattern recognition, automatic surveillance.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080104, Computer Vision, 50%
ANZSRC code: 080106, Image Processing, 25%
ANZSRC code: 080109, Pattern Recognition and Data Mining, 25%

Fields of Research (FoR) Classification

FoR code: 0801, Artificial Intelligence and Image Processing, 100%

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Over the last decades intelligent video surveillance systems have become increasingly important, due to the huge number of cameras used for security and surveillance of public areas. According to a recent report from the British Security Industry Association (BSIA), there are 4 to 5.9 million CCTV surveillance cameras in the UK alone. Monitoring this huge amount of information by direct human force is getting impractical. Hence security demands have been arguably the most important driver of research in advanced vision systems. Reliable automatic processing of the collected video information is a first step to initiate appropriate actions or to aid human agents in monitoring centers. Visual tracking is one of the core technologies in automatic video surveillance. An automatic visual tracking system is supposed to consistently locate the targets of interest throughout the video. In addition to surveillance systems, many other computer vision problems rely on visual tracking, including automated traffic monitoring, human-computer interaction, robotics and active camera systems.

However robust and reliable visual tracking is a very challenging problem in real world situations. Some of the well known challenges of visual tracking which may cause the system to fail in practice, include but are not limited to:

- Appearance variation of the targets due to pose changes, shadows, lighting variations, clothing, ...
- background clutter, noise and blur which may negatively affect the qual-

ity of the image and targets' appearance.

- Occlusions caused by the scene objects or other targets which affects the perfect visibility of the target seen from the camera point of view.

Handling occlusion situations in visual tracking is a challenging problem within uncontrolled environments. Tracking separated objects in the scene which are not interacting and occluding each other, does not seem very challenging with todays computer vision advancements. However long term real world scenarios always contain some sort of occlusion, like inter-object occlusion, self-occlusion or occlusion by the static scene occluders. Reliable tracking of the interacting targets would be more difficult, when they have similar appearance characteristics. Occlusion handling would be significantly more difficult when several targets in the scene cause frequent interactions and occlusions. 'Occlusion' and 'Confusion' (similarity between objects) have been recently classified to be among the most difficult challenges of visual tracking [CS10]. Hence we strongly believe that a proper occlusion reasoning strategy will largely improve the tracking performance.

One approach to resolve the occlusion problem, is to avoid it by placing cameras overhead, looking down on the plane of the moving targets [GSRL98, TSK00]. Another major trend for resolving the occlusion situations and maintaining the objects' identity during their interactions in the scene, is to use multiple cameras in order to resolve an occluded target by its other visible views [CGO00, DT01, DT00, Bat04]. However the focus of this thesis is on video surveillance with single cameras. Occlusion problem becomes increasingly important in video surveillance applications with single cameras, specifically when the camera is looking from a side view.

In visual tracking, occlusion events are a major cause for distractions and loss of the target tracks. Occlusions occur when distant targets are concealed by objects closer to the observer. Since an occluded object is not completely visible, unpredicted uncertainty in localizing the target is expectable. A common approach to handle occlusion events in visual tracking systems, is to suppress the errors introduced by occlusions to the optimization problem. Such approaches may blindly improve the robustness of the system in short term partial occlusions. However, they are not aware about the targets' visibility

status. Hence failures in full occlusions or long-term partial occlusions are very probable, due to gradual drifts and contamination of the target models. Various occlusion handling methods have been proposed in the literature, in accordance with different tracking frameworks. In this thesis, occlusion problem is studied for two types of visual tracking approaches separately, due to their different natures:

- Template based tracking, with an online template updating scheme;
- Tracking by detection based on off-line trained detectors, followed by a data association method;

Template-based trackers work by matching an appearance model of the target to the most likely location in consecutive frames. The tracking performance in such systems, is very dependent on the validity of the target model. Consequently, an updating strategy is required to maintain an up-to-date target model along the process. The updating process is crucial for template-based trackers, to enable them cope with variations of the target appearance in long term videos. However the online updating mechanisms, make the trackers vulnerable against template corruption, if being updated with improper candidates. Corruption of the target model is very probable in occlusion situation and may lead to permanent loss of the target.

On the other hand, tracking by detection based on off-line trained detectors, is a major trend in the context of multiple target tracking. Such trackers rely on off-line trained object models for detecting a category of objects, rather than the appearance model of individual targets. Consequently there is no online evolving target model in such tracking systems and thus no concern for template corruption. However, the main problem with such trackers is the general detection errors, such as missed detections and false alarms. Data association methods take advantage of temporal information for resolving the ambiguities and failures, by exploring correspondence among the detection results in consecutive frames. Missed detections are more probable in occlusion situations, due to inefficiency of the full body models for detecting the occluded objects. Furthermore the Non-Max-Suppression (NMS) process tend to suppress the nearby detections, which increase the confusion in occlusions and may lead to loosing the target tracks. The occlusion

problem in this context, may be addressed in any of the tracking stages: 1. Detection or 2. Data association. Occlusion handling in detection stage is equivalent to building stronger detectors, which perform better in occlusion situations. Despite various proposals for improving object detection in partial occlusions, the general detection errors are unavoidable. Hence the data association methods are supposed to handle the unresolved errors from the detection stage, while resolving the occlusion ambiguities simultaneously.

In this thesis we present two different occlusion handling frameworks for the two types of visual tracking systems, discussed above. An occlusion handling framework proposed for template based tracking systems, provides some tools for protecting the target model against corruption in occlusion events. The proposed tools improve the tracking performance in occlusion situation. Another occlusion handling method is developed for detection based trackers, within a multiple target tracking system. The proposed methods are able to handle occlusions caused by the static scene objects, as well as the inter-object occlusions. Self-occlusions are resolved to some extent, due to intrinsic robustness of the proposed systems against appearance changes. The proposed systems are applied for the purpose of pedestrian detection and tracking. However the suggested approaches are potentially applicable to other video surveillance areas, such as vehicles, animals, *etc*. Experiments and evaluations are conducted on various publicly available video surveillance datasets of pedestrians, which include challenging occlusion scenarios.

1.2 Main Contributions

The main objective of this thesis is to improve pedestrian detection and tracking in video surveillance, by providing some tools for occlusion handling. The proposed frameworks are evaluated on publicly available crowd surveillance videos. However, the provided tools are not specifically designed for human targets and are potentially applicable to general visual tracking. We propose two different occlusion handling approaches, for two typical tracking systems:

- Template based tracking systems ;

- Detection based tracking systems ;

The former proposal, incorporates multi-channel Spatio-temporal features for occlusion handling. In order to estimate the possibility of incorporating 3D multi-channel oriented features in real-time systems, we have developed a general purpose GPU engine, for evaluating the performance acceleration of a typical multi-channel filter set. The acceleration problem and the relevant GPU based parallel processing issues are discussed in a separate chapter of the thesis.

A brief overview of the main thesis contributions are presented in the following subsections. Further details of the proposed frameworks and contributions are provided in the next chapters, attributed to each of the topics.

Occlusion handling in template-based tracking

- An occlusion analysis framework is developed which empowers the tracking system to compete with state-of-the-art algorithms. Occlusion detection in the proposed system considerably improves the template updating mechanism, towards maintaining a valid up-to-date target model. This leads to substantial improvement of the tracking performance in occlusion situation.

- The proposed framework incorporates a 'Bayesian model' based on Spatio-temporal Oriented Energy features to determine state of the target in the course of tracking and discriminate between 'Partial' and 'Full' occlusions. This is very helpful in video surveillance scenarios of public areas, due to frequent short-term occlusions. An adaptive template update mechanism based on this model, helps to prevent template corruption in occlusion and drift situations.

- Detection of the Full Occlusion state, provides required ground for the system to change the tracking strategy, when there is no visible target in the scene to track. Such situations can easily lead to distraction, if the tracking process is not aware of the occlusion event. Hence a tracker could blindly match a new location to the target, regardless of its visibility status and end up with a target loss.

- The occlusion model involved in the framework is used to generate an occlusion mask, which determines the visible target pixels. Hence the tracking performance is improved in occlusion events, by concentrating on visible parts of the targets.
- Qualitative and quantitative evaluations demonstrate the strength of the proposed framework against alternative strong trackers such as IVT [RLLY08], MIL [BYB11], SOE [CGW10] and sparse trackers L1-IVT [JLY12] and L1-APG [BWLJ12]. These evaluations highlight the effectiveness of the proposed system in occlusion situation.

Occlusion handling in pedestrian detection and tracking

- An efficient spatio-temporal clustering framework is designed to improve Non-Max-Suppression (NMS) in occlusion events and compensate for the general detection errors, like false alarms and missed detections. The clustering cost function entails temporal consistency in the motion and scale of the tracked targets across consecutive frames. The formalization of the spatial and temporal terms in the proposed cost function, provides a light-weight closed form solution for the problem which can be solved in real-time with a general CPU.
- The clustering framework is based on a bounded cost function, which leaves some non-associated members per frame. An instantiation method is proposed to establish new clusters for the emergent targets through monitoring the non-associated members. We propose to apply the standard NMS on non-associated members and combine the NMS results with a specific confidence score, to increase the reliability of the instantiated clusters. The confidence score is composed of a notion of 'Detection Frequency' and a 'Depth/Height' probability.
- A track consolidation method is developed to post-process the spatio-temporal clustering results and remove low confidence tracks. we propose to utilize the scene 'Entry/Exit' information, along with a track confidence score for consolidation. The track confidence score is estimated based on an 'Occlusion Matrix' among the clusters per frame, as well as a 'Depth/Height' confidence of the clusters along the track.

- The proposed framework relies on a full-body pedestrian model, which does not require high-resolution details of the body parts. Hence the system is compatible with the existing infrastructures and low resolution cameras. This feature along with the real-time performance of the system, make it attractive for practical video surveillance.

1.3 Dissertation Overview

This dissertation is comprised of six chapters. Following this introduction, an overall literature review provides a proper context for the whole thesis, by describing the research gaps in the current literature. However, all the thesis chapters are self-contained, with their own literature review and experiment section. The six chapters of the thesis, including the current one, are summarized below:

- **Chapter 1: Introduction.** In this chapter the importance of the occlusion handling problem in video surveillance and visual tracking has been motivated and the research problems addressed by this thesis have been briefly introduced.
- **Chapter 2: An Overview on Visual Tracking and Occlusion Handling.** Chapter 2 provides a literature review on visual tracking and occlusion handling approaches that are relevant to this dissertation. The research gaps and shortcomings that exist in the current literature is elaborated in more details to provide the proper context for the thesis.
- **Chapter 3: Occlusion Handling in Template Tracking Systems.** Chapter 3 proposes an occlusion analysis framework, based on motion dynamics of the targets, for the purpose of robust template tracking in video surveillance applications. The proposed system takes advantage of the multi-channel 'Spatio-temporal Oriented Energy' (SOE) features for representing the targets' dynamics. Hence some background preliminaries on SOE features are initially introduced in this chapter. We demonstrate that protecting the target model against corruption and maintaining a valid up-to-date target model in challenging real world

scenarios, empowers the system to compete with state-of-the-art trackers.

- **Chapter 4: Occlusion Handling in Pedestrian Detection and Tracking.** Chapter 4 introduces an occlusion handling approach for multiple pedestrian detection and tracking based on a spatio-temporal clustering method. The proposed framework performs in real-time and demonstrates a high capability for occlusion handling within a low resolution context.

- **Chapter 5: Hardware Acceleration, Parallel Processing.** Chapter 5 presents a GPU-based hardware accelerator for extracting the biologically inspired multi-channel Gabor features. The GPU kernel imitates the parallel structure of the initial visual cortex layers, composed of 'Simple' and 'Complex' cells. The Simple cells perform the main feature extraction, while the Complex units aggregate the features at common orientation and provide abstract information with local invariance, which describe the orientational structure of the input image. The empirical speedup gain of the GPU engine, indicates that the multi-channel Spatiotemporal Oriented Energy features, applied in preceding chapters, can be generated in real-time for VGA size frames.

- **Chapter 6: Summary and Conclusions.** Chapter 6 provides a summary on the main contributions of the thesis. Furthermore the possible future directions and improvements for subsequent investigations are discussed.

Chapter 2

An Overview on Visual Tracking and Occlusion Handling

2.1 Introduction

Visual tracking may be considered as a primary step in many video surveillance applications, such as abnormality detection, action recognition, object identification and object recognition. Other applications of visual tracking may be seen in human-computer interaction, such as the Kinect input device developed by Microsoft [SFC+11] and active camera systems. An active camera is supposed to automatically detect the moving targets, as they appear in the field of view and keep their track through a smooth camera motion [MBM+95]).

In order to put our research into a proper context, in this chapter we briefly review some representative algorithms in the field of visual tracking with single cameras. The methods discussed in this chapter, are parts of the main stream in visual tracking with single camera, which have attained a widespread attention in the research community. In the last decade, these algorithms have seen significant advancements while being applied as basic building blocks in various visual tracking systems. Hence these major trends of monocular visual tracking, deserve a special attention in the context of this thesis. More comprehensive reviews of visual tracking systems are available in the computer vision literature [YJS06, Can08]. In the second part of this chapter (*c.f*. section 2.3), we will shortly review the proposed approaches for

occlusion handling in single camera visual tracking.

2.2 Visual Tracking Systems

Historically, tracking systems were initially introduced for following the point targets in RADAR systems. In a RADAR system, the discrete points on the screen represent the flying objects like airplanes. A major trend of statistical solutions for the problem of discrete point tracking in RADAR systems emerged between 1960's to 1980's [Kal60, Sit64, SS71]. However the more recent deterministic point trackers in computer vision domain between 1980's to 2000, relied on more basic assumptions such as point proximity and constant velocity. The main challenge in discrete point trackers is about following numerous targets with identical appearance. Thus the only cues to seek correspondence between point targets in consecutive frames are their position and velocity information. A typical approach in the deterministic point trackers is to perform an inter-frame "'proximity"' and "'velocity"' based matching between the points on a bunch of adjacent frames [SJ87]. Some extensions expanded the point tracking systems with occlusion handling and resolving entry/exit of the targets over time [SS90, RS91, SS03].

Computer vision as the science of automatic image and video processing, started to flourish with the great advancements of computer technologies. Hence, more detailed object descriptions and complex tracking algorithms, were tractable following their initial introduction. The main stream of visual tracking in computer vision has been established around some key tracking algorithms. In this section we discuss some of the representative visual tracking approaches, along with part of their evolutions in the past decade.

Bayesian methods

The main algorithms among the stochastic methods are 'Kalman filter' and 'particle filters', also known as sequential Monte Carlo. These methods are mathematical tools with diverse applications in many fields including signal processing and control systems. Visual tracking in this context is cast as a classic stochastic problem, where the statistically optimal state of the target (i.e. position, scale, orientation, ...) in each frame, is estimated based on the

observation variables (observed image/features) and state transition model, through a recursive Bayesian framework.

Kernel-based object tracking

This category of trackers employ a histogram representation of some target features (such as color, intensity, oriented gradient, ...), as the target model for visual tracking. A histogram model inherently destroys the spatial distribution of the target features and aggregates them all over the target support. Application of kernel histograms in visual tracking was popularized with the mean shift tracker introduced in 2000 by Comaniciu *et al.* [CRM00]. The mean shift tracker constructs the target template histogram in the first frame, which is used for tracking throughout the rest of video sequence. Iterations of the mean shift algorithm estimates a target candidate which is an optimal match for the target model. The Bhattacharyya coefficient was proposed as the similarity metric for finding the best match at each frame [CRM00]. For creating the histogram template, a kernel function is also applied to weight the pixels in the model based on their distance to the target center. Thus the closer pixels to the target center have a higher influence in constructing the histogram model, due to their higher reliability for representing the target. In other words, off the center pixels are more likely to be related to the background or other occluding objects in the scene. With this approach the proposed tracker gains some level of robustness to partial occlusions.

In a proceeding work, Comaniciu *et al.* [CRM03] proposed to utilize Kalman filter for predicting the target position in the next frame. This prediction was used as the initial state for the mean shift iterations at each frame, to help the mean shift tracker converge faster. Various extensions to the original mean shift tracker have been suggested in the literature, including the proposals for incorporating some level of spatial information in the histogram representation by subdividing the target into cells and assigning a histogram to each cell [NSHY08, ARS06].

Kanade-Lucas-Tomasi (KLT) tracking

Motion estimation based on the original image alignment algorithm of Lucas-Kanade [LK81], has been a major trend in visual tracking systems since its original introduction in 1981. In this approach, the motion estimation technique is based on the brightness constancy constraint, which implies that the intensity of the target pixels remain constant while they move throughout the video frame. This simple criterion which has been the basic foundation for optical flow estimation [BB95] and feature tracking [TK91, ST94], opened a new horizon in computer vision since its initial introduction in early 1980s. The constancy constraint can be practically applied to any features of the target pixels instead of intensities for motion estimation, i.e. the criterion may be expressed more generally in this way: "'The features of the target pixels remain constant while it is moving through the image frame"' ($F(\mathbf{x}, t-1) = F(\mathbf{x}+\vec{u}, t)$, where $\vec{u} = (u_x, u_y)$ and $\mathbf{x} = (x, y)$).

Ideally this equation should hold for every single pixel. However due to the mathematical approximations in derivation of the optical flow constraint, as well as the noise and measurement errors, there is some non-zero error in $F(\mathbf{x}, t-1) - F(\mathbf{x}+\vec{u}, t)$. Consequently motion estimation turns out to a minimization problem over the pixels within a local neighborhood. On this line another class of trackers gradually emerged, with a primary component which exploit the pixels' features (mostly intensity), in order to estimate the inter-frame motion of the target. In this framework tracking is cast as a minimization problem over the target support throughout the tracking process. If the target is assumed as rigid, like the case of the original work of Locus and Kanade [], then the the target motion is simply modeled by a translation vector which describes the horizontal and vertical displacement of a target region between two sequential frames. The estimated motion vector matches a reference image region (target of interest), to the optimal location in the next frame.

Inspired by this work, Bergen *et al.* [BAHH92] developed a model-based motion estimation framework in 1992, that accommodates more complex motions in addition to pure translation, such as rotation(2D or 3D), scaling and shearing. This framework provides the foundation for estimating

nonuniform motions of the pixels inside the target support. Thus the target deformations which occur due to change of pose throughout the video frames, was modeled to some extent in this framework. The minimization problem in this model-based motion estimation framework, will be:

$\underset{k}{\operatorname{argmin}} \sum_{\mathbf{x}} g(F(\mathbf{x}, t-1) - F(\mathbf{x} + \vec{u}(\mathbf{x}; k), t))$,

where F is the pixel's feature and motion vector $\vec{u}$ is defined as a parametric function of the pixel location with parameter vector k.

In this tracking framework, a target template is captured in the first frame of the video from the specific target location ($T(x, y) = F(x, y, t0)$) with the defined features F (intensity or ...). The tracker performs parametric matching of the target template to the image sequence to find out optimal candidate region in each frame. The estimated affine motion parameters, define the target's motion among the image sequence and forms a trajectory. We have utilized a similar KLT based tracking framework for part of our research, which will be discussed further in Chapter 3. We address this category of trackers as pixel based trackers in this dissertation.

Eigen Tracking

Eigen tracking initially emerged based on the idea of using lower-dimension subspace representations (Eigen-spaces) for modeling the target, as it appeared in works of Hager *et al.* [HB96] and Black *et al.* [BJ98] in late 1990s. A 'Principal Component Analysis' (PCA) was applied to calculate an appearance-based object representation in Eigen-spaces. They demonstrated that image variations due to illumination and pose change could be modeled in low-dimensional subspace [HB96, BK96, BJ98]. Subsequently they proposed tracking systems based on this Eigen representation which was shown to be robust against illumination and pose variations. However the initial Eigen trackers, required pre-training with a large set of training images (containing the whole range of appearance variation), in order to construct the eigen-basis. As a result the target model was fixed during the tracking and could only handle the pre-trained cases.

Ten years later, the eigen trackers started to popularize with the Eigen tracking system proposed by Ross *et al.* [RLLY08], which didn't require a

fixed pre-trained Eigen basis model prior to tracking. The proposed Eigen tracker instantiated the target model in the first video frames and incrementally updated the Eigen-basis on the fly. This was achieved by utilizing an incremental principal component analysis (PCA) algorithm to adapt the holistic appearance model to lighting and pose variations in the course of tracking. Eliminating the pre-training stage and proposing an online evolving model which could adapt to considerable appearance changes, was an important improvement which made the Eigen trackers suitable for practical visual tracking. Encouraged by this main contribution, they named their system as Incremental Visual Tracker (IVT).

Sparse Tracking

Inspired by advancements in sparse representation and its applications in signal processing and computer vision (*e.g.,* background subtraction [CSD+08], face recognition [WYG+09], ...), Mei *et al.* introduced the first visual tracking system based on sparse representation in 2009 [ML09]. This tracker models the target appearance sparsely through a set of representative templates, which are dynamically captured from the most reliable candidates throughout the tracking process. Moreover, noise and corruption, occlusion and changes in background are directly modeled by means of the positive/negative trivial templates. In the proposed framework by Mei *et al.*, visual tracking is cast as a sparse approximation problem in a particle filter framework, which is solved by ℓ_1 minimization formulation. More specifically target candidates in every new frame are represented by a sparse linear combination of the basis target templates and trivial templates. Then the optimal candidate with minimum error demonstrates the new target location and the tracking continues by propagating the sample distributions within the particle filter framework. Following the optimization method applied for the problem, sparse trackers are sometimes referred to as ℓ_1 trackers in the literature.

Other variants and applications of sparse trackers have been proposed afterward [MLW+11, ML11, BWLJ12, JLY12]. For example, a 'Structural Local Sparse Appearance Model' was introduced by Jia *et al.* [JLY12] which exploits the strength of both sparse representation (ℓ_1 tracker) and the incre-

mental subspace learning (IVT tracker). The so-called IVT-ℓ_1 tracker adapts its positive/negative templates to the changes in target's appearance with an incremental updating strategy. To this end, a linear combination of the PCA basis vectors and the ℓ_1 templates are utilized for modeling the estimated target.

Tracking by Classification

Some recent developments in visual tracking suggest improved performance can be achieved if the tracker is coupled with a dynamic detector/classifier component. In this context visual tracking is cast as a classification problem, in which a detector (static/dynamic) exploits the dissimilarity between the object and background to localize the target of interest in consecutive frames [Avi04, Avi07, BYB11, KMM12]. The general trend in most of the related studies, is to build a target appearance model in the first frame and update the model during the process using machine learning techniques. Some studies suggest to start with a target model which is trained offline and evolve the model during the tracking process. Such approaches are mostly applied in single target tracking.

One of the initial tracking systems in this line was proposed by Avidan (2001) [Avi01, Avi04] who suggested to use an off-line trained SVM classifier in a mathematical framework which resembles the standard optical flow equations. In this framework maximizing the SVM classification score is taken as the optimization criterion instead of the optical flow constancy constraint for computing the image gradients and locating the target. Hence rather than performing the calculations on two successive frames, like conventional optical flow systems, the support vectors take the role of the second image. This implies that the new frame is matched against the patterns used for training the classifier and calculations are performed on single frames. Although this algorithm has demonstrated success in vehicle tracking, but the static SVM classifier requires a lot of effort for pre-training with thousands of images of vehicles and non-vehicles in various situations.

Avidan proposed another tracker based on binary classification in [Avi05, Avi07]. A set of weak classifiers are trained online and combined into a strong classifier using Adaboost. This classifier is used to discriminate be-

tween the target pixels and the background pixels. The target of interest is then located by estimating the peak of the confidence map calculated for target pixels. With every new detection, new weak classifiers are trained and added to the ensemble of weak classifiers, to adaptively update the target and background model. Collins *et al.* [CL03] also treated tracking as a binary classification, in which the most discriminative RGB space is identified from a set of different color features. The selected feature space is then used for building a confidence map to discriminate the target from background and estimate the target location. Several variants of online learning boosted classifiers have been proposed in the literature, such as the asymmetric boosted classifiers by Pham *et al.* [PC07, PHC08], which are beyond the scope of this study.

A major challenge in the typical adaptive classifiers, as mentioned above, is that inaccurate tracking may lead to introduction of incorrect samples to the target's appearance model. This can gradually corrupt the target model and cause failure in long-term tracking. As this is a general problem in visual tracking, there has been various studies around it and some attempts to ameliorate the drifting and template corruption to some extent. However there is no general solution to this challenging problem yet. To improve on this issue, Babenko *et al.* [BYB11] proposed to generate a bag of templates from blocks around the current estimation of the target, rather than picking only one candidate for updating the model. Then a Multiple Instance Learning (MIL) framework was proposed for object tracking and learning a generative target model based on boosting weak classifiers into a strong classifier. In this approach the candidates are collected in bags and labels are assigned to the bags rather than individual samples. A positive bag normally contains a few bounding boxes around each object. By definition, a positive bag is assumed to have at least one positive sample, otherwise it is considered as a negative bag. Using bag of positive and negative templates can potentially improve the occlusion handling and drifting problem. However the ambiguity for choosing the most proper sample from the bag of positive templates, is passed to the learning algorithm. On the other hand training new weak classifiers with all the samples inside a positive bag, can potentially degrade the target model, since some of the samples inside the bag may be incorrect

candidates.

Kalal *et al.* [KMM12] propose a long-term tracking system by combining a tracker and a dynamic detector which bootstraps binary classifiers with structural constraints. In order to use more reliable candidates for online training of the classifiers and object model, a pair of 'P-N' experts are introduced to identify the detection errors at each iteration and use them for updating the model. P-expert is designed based on temporal structure (smooth motion) to identify false negatives and add them to the training set with positive labels. While N-expert exploits the spatial structure to identify the false positives, to be used as negative training samples. The dynamic classifier is updated by using this labeled training set, to avoid similar errors in future frames. With this approach drifting problem is alleviated to help long-term tracking.

Tracking by Detection

The visual tracking approaches discussed so far are generally adapted to single target tracking systems. A more recent major trend in multiple target tracking, is based on an object detection stage followed by data association. Two main approaches for target detection in this context are:

- Detection of foreground moving objects via background subtraction ;
- Detection of foreground objects with offline trained object detectors ;

Following the initial detection stage, data association among detected regions is supposed to merge the detection results across frames towards establishing target tracks.

Senior *et al.* [SHT+06] proposed a visual tracking system based on background subtraction and blob detection. The foreground blobs in consecutive frames are associated to form the target tracks. An appearance model is constructed on every blob which helps to improve the targets localization during tracking. The generated target models are also used to solve track correspondence in data association and resolve ambiguities in occlusions.

Andriluka *et al.* [ARS08] propose a part-based pedestrian detector within a pictorial structure model for the first stage of the tracking system. The offline trained detector is used to detect pedestrians in single frames. Temporal coherency among the detection results are then exploited within a three

stage probabilistic data association scheme in short, middle and long periods. A generative appearance model is extracted from short term tracklets, which is utilized along with a dynamical model to construct long term tracks of the targets.

Andriyenko *et al.* [AS11, MRS14] use a sliding-window linear SVM detector based on HOG [DT05] and HOF [WMSS10] features for pedestrian detection. Then a non-convex cost function is introduced for data association among the detection results in consecutive frames towards estimating smooth target tracks. Appearance model of the targets along with an occlusion model and a motion model are also incorporated in the proposed cost function.

2.3 Occlusion Handling Approaches

Among various well known challenges of visual tracking discussed earlier in section 1.1, occlusion problem is comparatively an under researched area in video surveillance. Various tracking frameworks have different requirements for handling occlusion situations due to their diverse natures. To deal with the problem, in the first step we need to understand different issues that occlusion may cause for a tracking system.

As mentioned previously 'Tracking by Detection' relying on off-line trained detectors, is a major recent trend in the context of multiple target tracking. However the template based trackers discussed in section 2.2 are generally adapted to single target tracking. Although some of them may be extended to multiple tracking systems as well. Different structures of the two mentioned tracking categories, naturally suggest distinct strategies for occlusion handling. In the following we briefly review the the occlusion handling challenges and the proposed solutions for the two tracking approaches, to put our research into a proper context. In the next two chapters we present two different frameworks for occlusion handling in template based tracking and detection based tracking.

Occlusion Handling in Template based tracking

Template-based trackers typically contain two main components: 1) an appearance model which defines the target by some features, 2) a search strategy for localizing the target in the frame through matching the target model to the most likely location. Tracking performance in such systems which rely on appearance model for tracking the target, is very dependent on the validity of the target template along the process. Trackers which employ static appearance models, generally learned on the first frame [BJ98, CRM00, ARS06], are not able to adapt with appearance changes and cannot operate in long-term. In real world video surveillance scenarios, the targets appearance are constantly changing among the video due to various issues such as non-rigidity of the targets, pose change, lighting variations, probable clothes change, *etc*. Hence maintaining an up-to-date target template is critically important for high performance tracking in long term. Due to variations over time, a template based tracking system is required to update its appearance model progressively and keep it up-to-date.

There are two main challenges for occlusion handling in this context: First, the occluded object which is not completely visible in the frame, may cause unpredicted uncertainty in localizing the target. Second, the adaptive schemes for maintaining up-to-data target model, are generally susceptible to be corrupted in occlusion or drift situations. In other words, updating the appearance model with occluded or inaccurate candidates captured from the current tracker location, leads to degrading the target model and drifting from the true target location.

Various approaches have been proposed in the literature to improve the tracking performance and robustness in occlusion situations. However tracking the occluded targets, specially in long periods and template contamination due to occlusions are still open research problems in visual tracking. For instance the multiple instance learning framework (MIL) in Babenko *et al.* [BYB11], demonstrates some level of success in partial occlusions. This may be a result of the multiple positive instances captured from around the target, which slightly incorporates the background in the model and can guide the tracker during the short-term partial occlusions. Some approaches propose to

divide the target into cells and model each cell separately [ARS06, CPB09]. Hence by tracking separate cells, an improved performance in partial occlusions is expected.

In the context of sparse tracking [ML09, ML11] noise, corruption and occlusion is modeled at pixel level by means of the positive and negative trivial templates, which can potentially improve the tracking performance in occlusion situations. Furthermore the sparse coefficients can be used to build an occlusion map, due to containing rich information about image corruptions and occlusions [MLW+11, BWLJ12].

Senior *et al.* [SHT+06] proposed an example of a multiple target tracking system based on background subtraction and blob detection, which maintains the appearance model of the tracked objects over time. The appearance models are utilized to resolve occlusion situations through a layering process which determines the depth ordering of the occluding targets. The appearance models also improve the target localization during the tracking process.

In spite of the improvements for occlusion handling in the tracking literature, a common problem among most of the trackers is their blind updating strategies. Hence the target models are susceptible to contamination during long-term occlusions, which may lead to drifts and failures. In Chapter 3 we propose an occlusion modeling framework based on Spatiotemporal Oriented Energies, which protects the target model against corruption in occlusions. We demonstrate that protection of the target model in template-based trackers, leads to significant improvement of tracking performance under challenging occlusion situations. A more detailed review of the existing occlusion handling approaches for model-based tracking systems are also provided in Chapter 3.

Occlusion Handling in Detection based Tracking

In this part we discuss the detection based tracking systems, which rely on off-line trained object detectors rather than the appearance models of the individual targets. Basically there is no template matching and evolving target templates in this context to be contaminated during an online updating process. Hence the occlusion problem in this context is inherently different compared to the template based systems, thus requiring new strategies for

handling the problem. The multiple-target trackers in the current discussion is composed of a general object detector followed by a data association scheme, which temporally combines the detection results for generating the tracks. Consequently occlusion inference may be performed in either stages:

- Occlusion handling in detection stage: developing stronger detectors which are more robust to partial occlusions ;
- Occlusion handling in data association stage: seeking improved data association schemes to resolve occlusion situations ;

It is well known that even for the strongest human detectors which stand on the top of state-of-the-art [DABP14, FGMR10], the detection performance significantly drops in occlusion situations [DWSP12], due to limited visibility of the occluded objects. Part-based detectors might be able to handle certain partial occlusions, given the target resolution provide sufficient information for detection [WU08]. However it has been demonstrated that even a state-of-the-art part-based human detector like 'Deformable Part Models - DPM' [FGMR10], starts to fail at low occlusion rates [TAS12].

One conventional approach for occlusion handling in the detection stage, is to apply multiple part detectors which are trained for specific body parts such as 'Head, Torso, Legs' [EESG10] or 'Right, Left, Bottom and Upper' body parts [WWRS11]. Subsequently all the detection results are combined within a specific framework to achieve an improved performance in partial occlusions. However due to the general lower performance of the part detectors compared to full-body detectors, achieving an acceptable performance in occlusion situations is questionable.

Some studies propose to build an occlusion map for the scene and apply the proper part detectors on visible areas of the targets [WHY09, GPK11, EESG10, WWRS11]. Enzweiler *et al.* [EESG10] calculate discontinuities in depth and motion of the objects based on dense stereo and optical flow to estimate the occlusion boundaries. While Wojek *et al.* [WWRS11] propose a light-weight approach to use the monocular scene geometry such as common ground plane and objects height, for estimating the depth and constructing the occlusion map. However the low performance of the part detectors is still an issue in the mentioned approaches.

Tang *et al.* [TAS12] propose a completely different approach for occlusion handling in pedestrian detection. They take advantage pf the pedestrian/pedestrian overlapping patterns for improving the detection performance in crowds, through training joint detectors to detect pair of occluding pedestrians. However scalability of this method and its generalization to other typical occlusions, like occlusions with the scene occluders or other moving objects, remain unsolved.

Undesirable errors of object detectors such as missed detections and false alarms are still inevitable in spite of all the progress in this field. Data association methods are supposed to resolve such faulty situations by taking advantage of temporal information. In other words, correspondence between the detections in consecutive frames are exploited to generate target tracks and resolve the detection ambiguities.

Some studies suggest to combine occlusion handling strategies within the data association framework. For instance, Milan(Andriyenko) *et al.* [AS11, MRS14] propose an energy cost function incorporating spatial and temporal terms to explore the data correspondence in consecutive frames. An occlusion model is integrated among the spatial terms to penalize existing targets with no evidence. However the proposed cost function is highly non-convex and computationally very intensive, which is far from a real-time performance.

Andriluka *et al.* [ARS08] propose a probabilistic approach for data association among the position, scale and articulation of the detected body parts. The proposed framework implements temporal coherency in three sequential stages. Coherency in the dynamics of the body limbs within a walking cycle are exploited to extract short tracklets. The tracklets form longer tracks between major occlusion events within a Hidden Markov model (HMM). Finally major occlusions are managed through association of estimated tracks using an appearance model and a coarse motion model. Furthermore the limbs dynamic model over a walking cycle, provides a means to somewhat handle partial occlusions in short term. Besides the computational complexity of the system, a sufficiently high resolution is required for such a detailed representation of the body parts.

Table 2.1: Summary of visual tracking paradigms regarding the occlusion problem;

Template based Tracking	Tracking target cells [ARS06, CPB09, KNHH11] Sparse tracking [MLW+11, BWLJ12, JLY12] Layering multiple targets with appearance [SHT+06] Super-pixel tracking [WLYY11]
Detection based Tracking (Occlusion addressed ... in Detection stage)	Combining multiple part detectors on visible areas: [WHY09, GPK11, EESG10, WWRS11] Occlusion map based on: SVM scores on target cells [WHY09] Discontinuities in stereo depth and motion [EESG10] Depth estimation via monocular scene geometry [WWRS11] Detectors trained for overlapping patterns [TAS12]
Detection based Tracking (Occlusion addressed ... in Data Association)	Energy function comprising occlusion model [AS11, MRS14] Probabilistic association via HMM [ARS08]

The above mentioned solutions for the occlusion problem in video-based detection and tracking are summarized in Table 2.1. More details on the occlusion handling approaches in the context of 'Tracking by Detection' is provided in chapter 4, along with a novel solution for the occlusion problem in this context. In chapter 4 we present a data association cost function, which explores temporal consistency in the motion and scale of the tracked targets. The proposed cost function has an efficient closed form solution, which provides a real-time performance. The proposed framework demonstrates a high capability for occlusion handling in a low resolution context, which makes it suitable for the existing infrastructures and low resolution cameras.

2.4 Summary

To recapitulate, the field of visual tracking is wide and diverse, with a history of more than 50 years. The literature review provided by this chapter is focused on monocular tracking with single cameras, to introduce a bigger picture of the family of trackers employed in the current thesis. The existing solutions to the problem of occlusion handling, are reviewed in two distinct family of tracking methods separately, to provide a greater context for understanding the main contributions of this thesis. Different aspects of the occlusion handling problem among the two distinct tracking approaches are

introduced and their shortcomings within the field of tracking are discussed. More specifically, the problems engaged with occlusion handling in 'Template based tracking' and 'Detection based tracking' are comparatively discussed, to establish the main research problems of this dissertation, which are addressed in the upcoming chapters.

Chapter 3

Occlusion Handling in Template Tracking Systems

3.1 Introduction

In this chapter, we introduce a framework to exploit motion dynamics of non-rigid human targets for the purpose of occlusion handling in template-based tracking systems. Although an extensive amount of research has been dedicated to overcome the well-known challenges of visual tracking such as illumination changes [Can08], comparatively little attention has been paid to robust schemes for template updating. Template corruption is the Achilles' heel in almost any tracking system and our motivation in this thesis is to address this under-researched area, especially for video surveillance technology.

Integrated analysis of videos in spatial and temporal domains is the de facto standard in various computer vision applications, since their initial introduction by Fahle and Poggio [FP81], Adelson and Bergen [AB85] and Heeger [Hee88]. The 'Spatiotemporal Oriented Energy (SOE)' feature set is an integrated and modern framework, proposed for analysis of dynamic patterns based on their constituent space-time orientation structure in the video. This framework has been successfully applied in various computer vision applications such as 'Dynamic Texture Recognition and Scene Understanding', 'Action Recognition' and 'Visual Tracking', to name a few [DW10, DSCW10, DW11, DW12, DLDW12, CGW10, CW14].

Modeling moving objects by SOE features for the purpose of visual tracking was first explored by Cannons *et al.* [CW07, CGW10, CW14]. For simplicity in this chapter, we refer to this category of visual tracking systems as SOE trackers. Though SOE trackers have been shown to yield good performance in certain realistic situations [CW07, CGW10, CW14], their intrinsic limitations may prove a hindrance in some others.

One example is the tailing effect in SOE features, which is an intrinsic consequence of temporal filtering. Basically the presence of the moving target in various locations of n successive frames, appears as an energy tail in the SOE features. This makes the blobs of spatiotemporal energies look larger than the actual target size. Therefore, the target bounding box oscillates around the moving target over time, even for normal and non-occluded situations, which could lead to incorrect or unstable target localization (*c.f*. experimental results of the surveillance video in [CGW10, Can10]). Obviously, larger intra-frame target movements produce longer energy tails and higher inaccuracies in localization.

Another problem appears in scenarios where multiple targets move together in the same direction. In such situations the targets' energy blobs may be confused with each other and cause tracking errors. This is mainly according to the similar motion dynamics of the moving targets and the tailing effect of the SOE features. Moreover, if the motion direction of the targets rapidly changes in the video, their SOE based template will be invalid afterwards, due to the slow nature of the updating scheme in [CGW10] and the fast evolution of motion dynamics.

Unlike previous studies that utilize SOE features as fundamental tracking cues, we propose an SOE-based framework for occlusion modeling and detection of tracking novelties. In this chapter novelty is defined as partial occlusion events when the occluding objects have different motion dynamics (*c.f*. §3.4.2). The proposed occlusion detection system is designed for non-rigid human targets and works based on the targets' motion dynamics which are characterized through their 'Spatio-temporal Oriented Energies'. We design a Bayesian state machine to discriminate between the 'Partial Occlusion' and 'Full Occlusion' states of the target. Hence template updating mechanism is more reliable, as a result of protecting the target model against corruption

in novelty situations. This leads to a significant performance improvement in tracking under challenging occlusion situations. Furthermore, in 'Full Occlusion' situations a different tracking strategy is sought due to invisibility of the target. An occlusion mask is also estimated based on SOE features to distinguish non-occluded parts of the target. Hence tracking optimization can be performed using the non-occluded target pixels. This approach enhances the tracking performance in novelty situations, while avoiding the inaccuracies caused by the confusion of similar SOE blobs in SOE trackers. The proposed framework is extensively evaluated on several publicly available surveillance datasets in short-term/long-term occlusions and compared against state-of-the-art trackers. It is worth mentioning that our proposed framework can be seamlessly fused with various tracking schemes, as a means to protect the target template or the bag of templates.

3.2 Related Works

Handling occlusion in modern visual tracking is not overlooked by any means but the theme adopted in many studies is not a dedicated approach. Generally speaking, occlusion can be handled implicitly or explicitly in a tracker. In implicit solutions, the tracker mostly relies on a discriminative template and usually is able to handle short term partial occlusions [RLLY08, KL11, BYB11, ARS06, CPB09, ML09, ML11, JLY12, LYY13, ZYSL13, WCXY12, BL12, HJZ+11]. On the contrary, in explicit solutions a dedicated mechanism is foreseen to handle and guide tracker during severe occlusion [MLW+11, KNHH11, WLYY11, BWLJ12]. In the following text, we briefly overview some examples from each category.

Ross *et al.* suggested an incremental visual tracker (IVT) that incrementally updates a low-dimensional subspace representation of the target [RLLY08]. This is achieved by utilizing an incremental principal component analysis (PCA) algorithm to adapt the holistic appearance model to lighting and pose variations in the course of tracking. Partial occlusions can be considered to some extent in IVT, by involving a forgetting factor which reduces the effect of new data in constructing the low-dimensional subspace. The larger the forgetting factor, the lower is the sensitivity to short-term occlusions. Hence

the system preserves the non-occluded representation for a short period of time. Babenko and Yang [BYB11] proposed to generate a bag of templates from blocks around the current estimation of target and multiple instance learning framework for object tracking (MIL-Track). MIL-Track shows robustness against various appearance changes and partial occlusions to some extent. This is due to capturing multiple positive instances around the target which slightly incorporates the target background in the appearance model. Consequently the surrounding blocks can guide the tracker during the short-term occlusions. Such approaches could result in decent performance for short-term occlusions, but obviously fails in more challenging long-term and heavy occlusions.

A few recent studies adopted sparse representation for visual tracking, where the target appearance is modeled sparsely through a set of templates [ML09, ZYSL13, LYY13, ML11, WCXY12, JLY12, BL12, HJZ+11]. In these so-called ℓ_1 trackers such as proposed in [ML09], noise and corruption, occlusion and changes in background are directly addressed by means of the positive/negative trivial templates and the ℓ_1 minimization formulation. After all, occlusion is modeled at pixel level in the ℓ_1 tracker and though it might be handled to some extent (as a result of sparse formulation) a dedicated mechanism for detecting occlusions was not foreseen. Hence, the tracker could still suffer from long-term occlusions, which could introduce occluded templates to the tracker and cause drifting. A 'Structural Local Sparse Appearance Model' is introduced in [JLY12] which exploits the strength of both sparse representation (ℓ_1 tracker) and the incremental subspace learning (IVT tracker). The so-called IVT-ℓ_1 tracker adapts its positive/negative templates to the changes in target's appearance with an incremental updating strategy. To this end, a linear combination of the PCA basis vectors and the ℓ_1 templates are utilized for modeling the estimated target. To avoid contaminating the target model with occluded frames, the template set is updated by reconstructed target images using only the PCA basis vectors rather than the raw estimated target. Consequently the IVT-ℓ_1 tracker can handle partial occlusion by reducing the template contamination.

To achieve robustness against partial occlusion, some approaches divide the target (and the background in some cases) into multiple fragments and

model each fragment separately in the joint feature-spatial space [ARS06, CPB09]. One could expect a level of robustness against occlusion by tracking the fragments separately instead of the whole target. Nevertheless, similar to ℓ_1 tracker the aforementioned methods do not exploit a dedicated machinery for detecting occlusion. In this text methods such as ℓ_1 tracker which show some level of robustness to occlusion without benefiting from a dedicated and explicit mechanism are dubbed implicit approaches. The common problem among all of the tracking schemes incorporating implicit approaches for handling occlusion, is that the appearance models are vulnerable to be contaminated during long-term occlusions due to their blind update strategies. Hence the tracker is susceptible to failures or drifts when a heavily occluded target is introduced into the template set.

Recently, a few general frameworks for explicit occlusion modeling have been introduced [MLW+11, KNHH11, WLYY11, BWLJ12]. The 'Bounded Particle Resampling' BPR-ℓ_1 tracker [MLW+11] detects occlusion by analyzing the sparse coefficients obtained from the ℓ_1 minimization problem. The sparse coefficients inherently contain rich information about image corruptions and occlusion and can be used to build a binary occlusion map. The small areas and holes in the binary map are removed by morphological operations. The occlusion is then detected by measuring the largest connected component in the occlusion map. In [BWLJ12] an improved real-time version of the BPR-ℓ_1 tracker is proposed based on the accelerated proximal gradient (APG) approach for solving the ℓ_1 minimization problem. The occlusion detection mechanism in APG-ℓ_1 is similar to BPR-ℓ_1, while its tracking accuracy is improved to some extent.

In [KNHH11] an occlusion detection technique is proposed through applying a classifier based on observation likelihoods of the target regular grid cells. In [WLYY11] a discriminative appearance model based on superpixels is utilized for tracking and a target-background confidence map is used for computing a maximum a posterior estimate of target location. The target confidence of the MAP estimate is then used as a clue to detect heavy or full occlusions, which in turn helps maintaining the target information in long-duration occlusions.

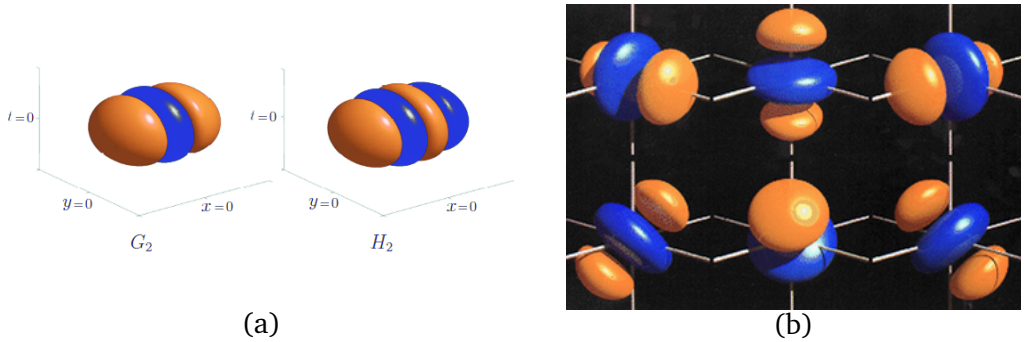
The proposed framework in this chapter, analyzes the targets' motion dy-

namics based on SOE features for the purpose of occlusion modeling and novelty detection. We start by introducing the 'Spatiotemporal Oriented Energies'.

3.3 Spatiotemporal Features

In this section we briefly review the basics of SOE features, which are fundamentally a decomposition of the video stream into spatiotemporal oriented energy planes. This serves as foundation for future development in this chapter, in definition of the targets motion model.

Decomposition of a video sequence into space-time oriented energy planes is performed by filtering the video sequence with a set of 3D band-pass filters in various space-time orientations and combining the resulting energy channels within a specific model. In this work, we will make use of the second derivative of 3D Gaussian filters $G_2(\theta)$ and their Hilbert transforms $H_2(\theta)$ for space-time filtering of videos, where θ specifies the 3D direction of the filter axis of symmetry. A sample quadrature pair filter along the x-axis is shown in figure 3.1-(a). The functions $G_2(\theta)$ and $H_2(\theta)$ along any orientation θ can be synthesized through a linear combination of a small set of basis functions [FA91, DG05]. For instance, figure 3.1-(b) demonstrates all the 3 dimensional basis functions of the second derivative of Gaussian $G_2(\theta)$ for any θ. However ten basis functions are required to span the space of functions $H_2(\theta)$ (Hilbert transform of $G_2(\theta)$) for all θ. Hence the quadrature pair filters (G_2, H_2) are broadly tunable, separable and steerable filters.



(a) (b)

Figure 3.1: (a) Iso-surface profile of quadrature pair filters (G_2, H_2) along the x axis, $\theta = (1, 0, 0)$. (b) 3D basis functions of Gaussian derivatives $G_2(\theta)$ for all θ.

Let $I(x, y, t)$ be the intensity of a point (x, y) at time t. Given θ, a measure of oriented spatio-temporal energy is obtained using:

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left[G_2(\vec{\boldsymbol{\theta}}) * I(x, y, t)\right]^2 + \left[H_2(\vec{\boldsymbol{\theta}}) * I(x, y, t)\right]^2 , \tag{3.1}$$

where $*$ denotes the convolution operator.

An efficient way of calculating Eqn. (3.1) for various space-time orientations is to first convolve the input sequence by the set of basis functions of G_2 and H_2. Then the filtered response along the orientation θ, *i.e.,* $E_{\boldsymbol{\theta}}$ can be obtained by a linear combination of the filtered basis images, where the coefficients of this linear combination are functions of the space-time orientation θ [DG05].

A 3D band-pass filter does not respond to space-time patterns oriented along its axis of symmetry. Instead it is most responsive to all space-time patterns orthogonal to the 3D filter orientation [Sim93]. To make it clear, figure 3.2 illustrates a moving one dimensional pattern $(x - t)$ and its frequency response. As shown in the figure, the power spectrum of the pattern lies on a line perpendicular to the motion direction. More generally the frequency response of a single oriented space-time(/3D) pattern, such as a rigid moving object, manifests itself in a plane perpendicular to the orientation of the moving object.

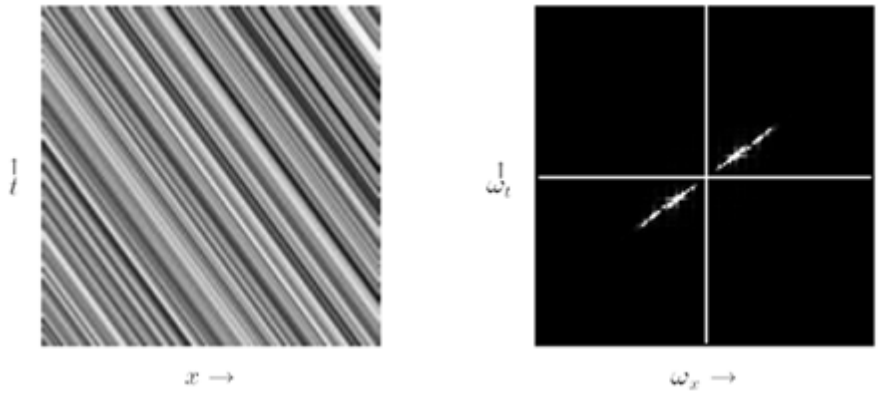


Figure 3.2: A moving one-dimensional pattern on left and the corresponding power spectrum on right (Courtesy of Simoncelli [Sim93]).

In order to span the frequency plane of the spatiotemporal patterns along specific space-time orientations, a set of at least $(N + 1)$ perpendicular Gaus-

sian derivative filters of order N are required [FA91]. Hence summation of the $(N+1)$ equally spaced SOE filters $E_{\boldsymbol{\theta}_i}$, extracts the dynamic energy along the perpendicular direction regardless of the local spatial structure [DW10] (*c.f.* figure 3.3):

$$\widetilde{E}_{\boldsymbol{n}}(\boldsymbol{x}) = \sum_{i=0}^{N} E_{\boldsymbol{\theta}_i}(\boldsymbol{x})\,, \tag{3.2}$$

where θ_i represents the $(N+1)$ equally spaced orientations orthogonal to motion direction $\boldsymbol{n}$ and each E_{θ_i} is calculated via the oriented energy filtering in Eqn. (3.1). Simoncelli [Sim93] demonstrated that summation of the ring of $(N+1)$ equally spaced filters in a plane, produce a smooth 'donut' shape, as illustrated in figure 3.3 and called it 'donut mechanism'.

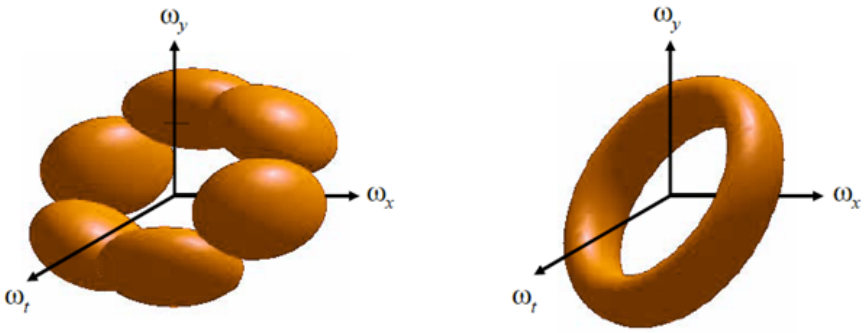


Figure 3.3: **Left:** A set of 3 equally spaced 2^{nd} derivative of Gaussian filters in a plane , **Right:** The corresponding level surface of the sum of power spectra of the filters, which is maximally responsive to the orthogonal spacetime orientation

For a plane with a unit normal vector $\vec{\boldsymbol{n}}$, the set of $(N+1)$ unit-length and equally spaced in-plane vectors are given as:

$$\vec{\boldsymbol{\theta}}_i = \cos\left(\frac{\pi i}{N+1}\right)\boldsymbol{\theta}_a(\vec{\boldsymbol{n}}) + \sin\left(\frac{\pi i}{N+1}\right)\boldsymbol{\theta}_b(\vec{\boldsymbol{n}}) \tag{3.3}$$

$$\forall\; i \in [0, N]\;,\; \boldsymbol{\theta}_a(\vec{\boldsymbol{n}}) = \frac{\vec{\boldsymbol{n}} \times \boldsymbol{e}_x}{\|\vec{\boldsymbol{n}} \times \boldsymbol{e}_x\|} \;\text{ and }\; \boldsymbol{\theta}_b(\vec{\boldsymbol{n}}) = \vec{\boldsymbol{n}} \times \boldsymbol{\theta}_a(\vec{\boldsymbol{n}})\,,$$

where $\boldsymbol{e}_x$ is the unit vector along the x axis. If u_x and u_y denote the velocities along x and y axis, $\vec{\boldsymbol{n}}$ (the space-time motion direction) is obtained as $\vec{\boldsymbol{n}} = (u_x, u_y, 1)/\|(u_x, u_y, 1)\|$.

In this work, we utilize $(N+1) = 3$ equally spaced 3D oriented filters

to measure space-time single orientation pattern or motion direction. The response of 3D oriented filters are combined into five space-time orientations, namely, leftward $L = (-1, 0, 1)/\sqrt{2}$, rightward $R = (1, 0, 1)/\sqrt{2}$, upward $U = (0, -1, 1)/\sqrt{2}$, downward $D = (0, 1, 1)/\sqrt{2}$, and the static/no motion channel $S = (0, 0, 1)$ along the time axis. The resulting energy measurements are dependent on local intensity contrast, while in our application just the relative contributions of space-time orientations are important. Therefore, a pixel-wise normalization of the energy measures is performed in the final step as follows:

$$E_{\boldsymbol{n}_i}(\boldsymbol{x}) = \frac{\widetilde{E}_{\boldsymbol{n}_i}(\boldsymbol{x})}{\varepsilon + \sum_{j=1}^{\Gamma} \widetilde{E}_{\boldsymbol{n}_j}(\boldsymbol{x})}\,, \tag{3.4}$$

where Γ is the number of applied space-time orientations ($\Gamma = 5$) and ε is a noise floor constant to avoid instabilities at points with small overall energy.

For visual purposes, figure 3.4 illustrates different channels of SOE features extracted from a synthetic video through the equation 3.4. The video sequence contains two square blocks which turn around the frame in parallel and opposite directions. A static structure is also placed in the center of the frame. Figure 3.4-(a) demonstrates the responses of SOE channels to horizontal motion of the blocks (rightward and leftward), as well as the static structure. Figure 3.4-(b) presents the responses of SOE channels to vertically moving blocks (upward and downward). We observe that every channel strongly responds only to the relevant motion dynamics, while the response to other motion directions are weaker.

Besides the aforementioned features, we need a notion of lack of structure (O) in this work. Eqn. (3.5) represents such a notion in video regions with insufficient information for estimating the flow:

$$E_O(\boldsymbol{x}) = \frac{\varepsilon}{\varepsilon + \sum_{j=1}^{\Gamma} \widetilde{E}_{\boldsymbol{n}_j}(\boldsymbol{x})}\,. \tag{3.5}$$

A significant amount of energy in E_O channel shows that the pixel neighborhood is devoid of sufficient structure to determine its motion dynamics. This implies that the value of dynamic energy channels $E_{\boldsymbol{n}_i}$ are not reliable for a voxel when E_O attains a high value. The features $E_{\boldsymbol{n}_i}$ along the five

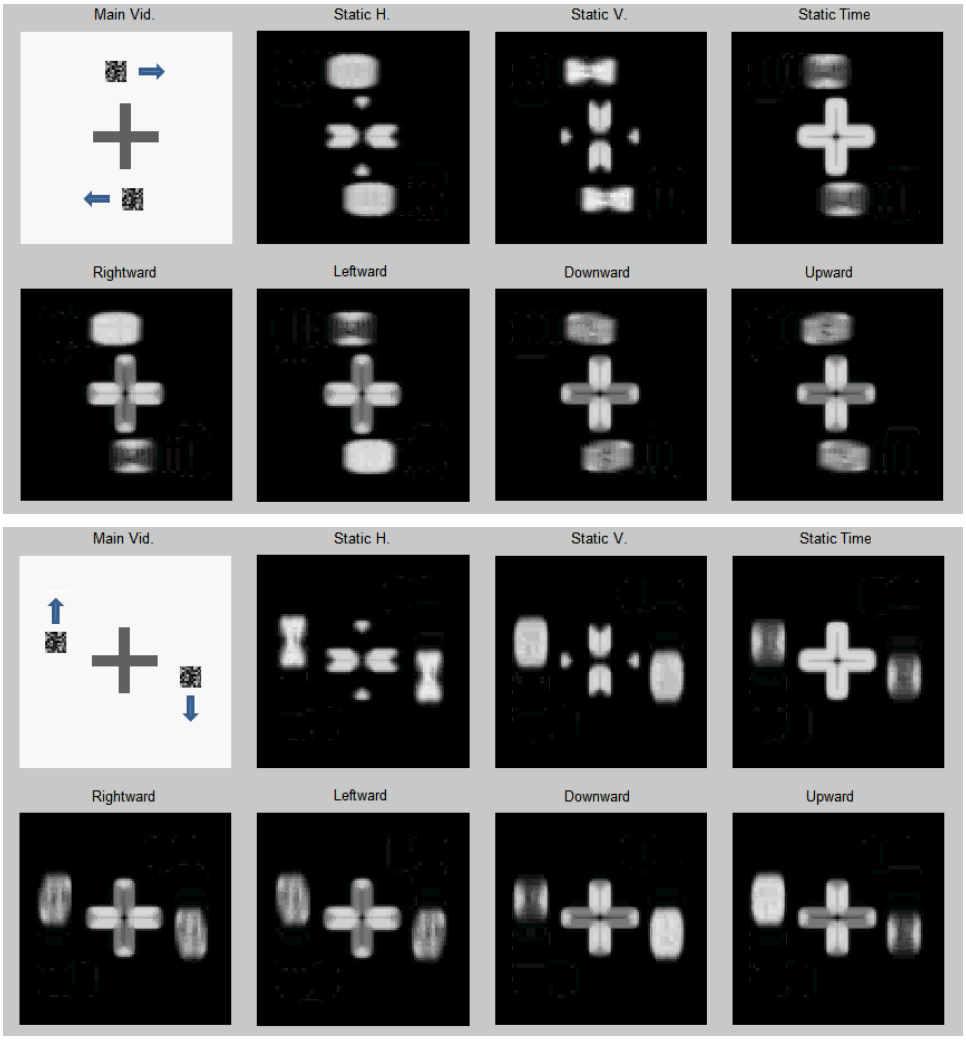


Figure 3.4: Responses of SOE feature channels to various motion dynamics: (a) Rightward, Leftward, Static. (b) Upward, Downward, Static.

spatiotemporal orientations and the unstructured channel provide a space-time description of the pixels' dynamics. We note that from Eqn. (3.4) and Eqn. (3.5):

$$\sum\nolimits_{\boldsymbol{n}_i} E_{\boldsymbol{n}_i} = 1 \quad , \quad \boldsymbol{n}_i \in \{S, R, L, U, D, O\} \, . \tag{3.6}$$

Following the original works in the literature [DW09, DW10], we use the term 'Marginalised SOE' or the abbreviated form 'MSOE', for the introduced features.

3.4 Technical Approach

We start this section by providing a high-level description of the components of our proposed tracking system, followed by a detailed explanation for each and every one of them. A conceptual diagram of our proposed tracking system is shown in Fig. 3.5.

The tracker module exploits the pixel intensities to calculate an average inter-frame motion vector for the whole target. Therefore, the motion of non-rigid target is modeled by one motion vector. The target of interest is described by two separate models, indicated by "Appearance" and "Motion Dynamics". The appearance model is employed by the pixel tracker for calculating the target motion vector. The motion model is exploited to estimate the dynamic status of the target as well as the novelty situation. The motion model is also utilized for calculating an "Occlusion Mask" which determines parts of the target window that most probably belong to the target of interest in novelty situation.

The "Novelty Detection" module prevents template updating in occlusions and novelties and serves as a mean to avoid template corruption and mitigate the drift problem. The "State-Machine" module estimates the state of the target in the course of tracking, to change the tracking strategy in occlusion situations. When the target is fully occluded and invisible, the system temporarily predicts the target location according to its most recent motion dynamics. The target models are adaptively updated in the course of tracking, while protected against corruption and non-desirable changes by means of the 'Novelty Detector'.

3.4.1 Tracking

In this work, we employ a customized version of the generic intensity based pixel tracker [BA96]. More specifically, let $M \in \{0,1\}^{w \times h}$ be a binary matrix of size $w \times h$ representing the occlusion mask (later will be defined) of the target $T_{w \times h}$. Moreover, let $\boldsymbol{u}(x, y; \boldsymbol{a})$ be the flow model of the target parametrized by the vector $\boldsymbol{a}$. Then the tracking problem is cast as a minimization problem to estimate an average motion for the whole target as

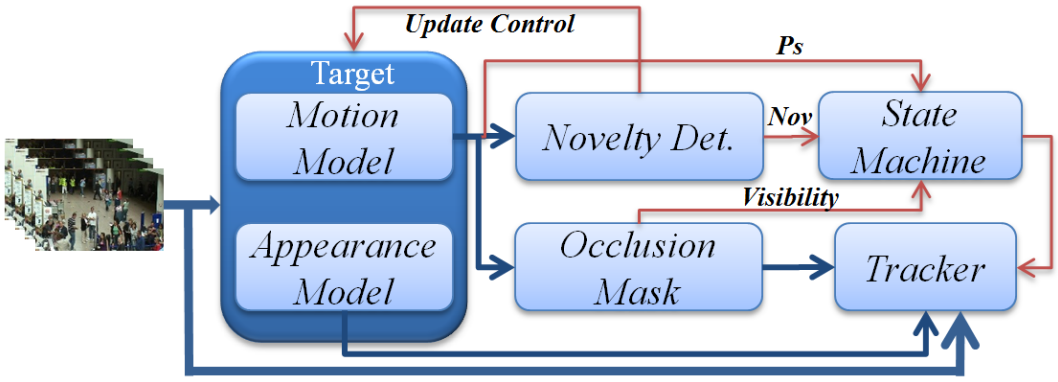


Figure 3.5: Block Diagram of Tracking System ; Red lines are Control Signals and Blue lines represent Data

follows:

$$\arg\min_{\boldsymbol{a}} E(\boldsymbol{a}) = \sum_{x,y \in T} M(x,y) \cdot \rho \left(\nabla I(x,y)^T \boldsymbol{u}(x,y;\boldsymbol{a}) + \frac{\partial I(x,y,t)}{\partial t}, \sigma \right), \tag{3.7}$$

in which $\nabla I = (\partial I(x,y,t)/\partial x, \partial I(x,y,t)/\partial y)^T$ and $\rho : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+$ is an error function with the scaling parameter σ. In this work we use the Geman-McClure robust estimator as error function (ρ) which not only diminishes the effect of outliers but also smooths out the effect of articulated object motions inside the target window [BA96].

$$\rho(x,\sigma) = \frac{x^2}{\sigma^2 + x^2}\,. \tag{3.8}$$

The binary occlusion mask M is defined based on relevancy of pixels to the target of interest. Consequently only highly probable target pixels with closely relevant "Motion Signatures" contribute to the optimization process.

In order to estimate the flow vectors within a closed form solution, the flow field can be modeled as a parametric function of the image coordinates. Common models of image flow to restrict motion vectors include constant, affine and quadratic. We use a constant velocity model in this work. However extension to more complex parametric models is straightforward [BA96]. The constant model $\boldsymbol{u}(x,y;\boldsymbol{a}) = (a_0, a_1)^T$ considers a constant average motion for all of the target pixels.

Optimization over the error function $E(\boldsymbol{a})$ is performed through a gradi-

ent descent procedure, *i.e.,* optimum model parameters which compose the motion vectors are estimated through a recursive equation:

$$\boldsymbol{a}^{(k+1)} = \boldsymbol{a}^{(k)} - \beta \frac{\partial E(\boldsymbol{a})}{\partial \boldsymbol{a}} \tag{3.9}$$

Optimization is accomplished in a hierarchical coarse-to-fine framework, to cope with the large target motions. Further details about the optimization process and parameters can be found in [BA96].

In "Full Occlusion" which is defined as the temporary disappearance of the object, the tracking strategy is changed to a short-term 'tracking by prediction'. In this situation, target location is estimated according to its recent average velocity. At the same time the tracker searches for the lost target energy blob with a similar 'Motion Signature' with respect to the last valid 'Motion Signature' of the lost target. The target is reported as lost, if the energy blob is not found within a predefined time period.

3.4.2 Occlusion Modeling

In this section we describe various components of our occlusion modeling system, which works based on the motion dynamics of the targets. We start by elaborating on the proposed motion model.

Target Motion Model

Normalized energies from Eqn. (3.4) and Eqn. (3.5) are used to determine the presence of particular spatiotemporal oriented structures related to specific motion directions. The coherent motion of target throughout the video frames corresponds to a single space-time orientation, which results in a dominant response in one component of the spatiotemporal oriented channels [DW09].

Target motion is modeled based on the extracted MSOE features of the pixels inside the target window $(E_S, E_R, E_L, E_U, E_D, E_O)$. According to Eqn. (3.6), scalar variables of the six-tuple MSOE features are dependent. Furthermore the pairs of scalar variables (E_R, E_L) and (E_U, E_D) are pairwise dependent. In other words, an increase of energy in one channel corresponds to a decrease of energy in the opposite channel, due to the single motion of pixels

at a time instant. Hence two new variables are defined which correspond to horizontal and vertical motion dynamics of the pixels:

$$E_H = (E_R - E_L)\ ;\ E_V = (E_U - E_D) \tag{3.10}$$

The triple $\{E_S, E_H, E_V\}$ is called the "Motion Signature" in this thesis. In object tracking, it is realistic to assume that most of the pixels of moving target demonstrate a coherent motion. This is especially true for rigid objects, though the outlier pixels such as the included background can cause some incoherencies. For non-rigid targets such as humans, more incoherencies among pixels might be expected. Nevertheless, it is still a valid assumption that for a joined moving object, the majority of target pixels behave coherently.

In light of the above discussion, the target MSOE feature variables are interpreted as random variables whose distributions are estimated and utilized for statistical modeling of the target motion dynamics. Due to the major concentration of the motion signature variables around a mean which is relevant to the target motion direction, a Gaussian distribution is considered for modeling. More specifically, the random variables $\{E_S, E_H, E_V\}$ are separately modeled by 1D Gaussian functions to construct the compound target motion model:

$$\{\ \mathcal{N}_S(\mu_S, \sigma_S)\ ;\ \mathcal{N}_H(\mu_H, \sigma_H)\ ;\ \mathcal{N}_V(\mu_V, \sigma_V)\ \} \tag{3.11}$$

where $(\mu_i, \sigma_i),\ i \in \{S, H, V\}$ are the mean and variance of the motion variables E_i over the target window, respectively. The variance of the Gaussian models are related to the nonrigidity level of the targets.

Target Status Determination

The various states of the target which can be segregated in the course of tracking include the non-occluded states 'Moving' and 'Stationary', as well as 'Partial' and 'Full' occlusions. In order to determine the target states, the system needs to evaluate the target status probabilities in non-occluded situations, while detecting the occlusion events simultaneously. This is performed

as elaborated in the following sections.

Status Probabilities

In non-occluded situations where the target of interest is completely visible, the mean of the Gaussian motion models represent the status of the target. Referring to Eqn. (3.4) and Eqn. (3.6), the Γ channel of normalised MSOE features E_{n_i} are in the range $(0, 1)$ and dependently fluctuate around $(1/\Gamma)$. Hence the oriented energy channels approach zero in non-structured regions, while E_O grows according to Eqn. (3.5). Consequently the energies in E_S, E_H and E_V are proportional to the status of the target, whether it is static or moving in a particular direction. For instance a static pixel demonstrates higher energy in E_S channel rather than horizontal (E_H) or vertical (E_V) energy channels. Using sigmoid functions, the energy values can be converted to pseudo-probability figures to represent the target status:

$$\begin{aligned} P_S &= \frac{1}{1+e^{-\beta(\mu_S-\alpha)}} \;;\; P_{hv} = \frac{1}{1+e^{-\beta(|\mu_H|-|\mu_V|)}} \\ P_H &= (1-P_S)\cdot P_{hv} \;\;;\;\; P_V = (1-P_S)\cdot(1-P_{hv}) \end{aligned} \tag{3.12}$$

where α is an energy threshold for 'Static' energy channel (E_S) to discriminate between 'Static' and 'non-Static' situations and β represents the slope of the Sigmoid curve in the transition area. The sharper the transition of Sigmoid curve, the closer is the system to binary decision making. The variable P_{hv} is defined as a transitional compound probability measure to simplify the declaration of P_H and P_V. We note that

$$P_S + P_H + P_V = 1, \tag{3.13}$$

which supports the intuition that the target 'Static' situation is mutually exclusive with its 'Horizontal' or 'Vertical' moving status.

Novelty Detection

Novelty in our problem is defined as partial occlusion events, where the occluding objects demonstrate different motion dynamics. In such situations

where more than one coherent motion is available in the target bounding box, the assumption for calculating Gaussian models of the target motion in 3.4.2 are violated. Consequently motion novelties can be detected by checking the validity of the Gaussian modeling, as in this situation motion random variables can not be modeled with single Gaussians.

To this end, we propose a straightforward approach to detect novelties. We note that two major motion dynamics in the target bounding box indicate the occlusion situation and the sample variance of the motion variables are affected in this case. Accordingly we propose the following metric to detect the novelty:

$$\sigma_{Nov} = P_S \cdot \sigma_S + P_H \cdot \sigma_H + P_V \cdot \sigma_V, \tag{3.14}$$

where $\{\sigma_S, \sigma_H, \sigma_V\}$ are sample variance of the motion variables inside the target bounding box, related to energy channels $\{S, H, V\}$. We note that the motion novelty may appear in more than one motion variable based on the status of occluding objects. For example when the target of interest and the occluding object are both moving horizontally, novelty appears just in the horizontal channel. While in the case that the target is moving diagonally, it obviously exhibits energy in both horizontal and vertical channels (P_H and P_V). Consequently depending on the status of the occluding object, novelty may appear in both channels. Since the coherent motion of the target is reflected in high energy channels, we expect to observe novelties mostly in the channels with high energies and probabilities.

In normal (non-occluded) situations, the variance of the Gaussian motion models represent the nonrigidity level of the target along with the effect of outliers, which varies for different targets and situations. Thus novelty can be detected only if a good estimate of the expected variance of the target motion model is available. For this purpose a simple model is trained online to adaptively estimate the expected variance of the motion random variables $\{R_S, R_H, R_V\}$ (which will be described in § 3.4.3). The R_i values are computed during non-occluded states to evaluate the normal nonrigidity level of the target in three motion channels. This way the difference between the

drift of σ_i in occlusion and non-occluded situation is coded. Based on the attributes R_i, an adaptive reference level (R_{Th}) is defined to represent normal nonrigidity of the target:

$$R_{Th} = P_S \cdot R_S + P_H \cdot R_H + P_V \cdot R_V \ . \tag{3.15}$$

Having R_{Th} as a reference for novelty detection at our disposal, the following figure is utilized to detect novelties:

$$F_{Nov} = g_u(\sigma_{Nov}/R_{Th} - \tau_{Nov}), \tag{3.16}$$

where g_u is a unit step function and τ_{Nov} is a threshold for estimating the logic decision about novelty detection. In all our experiments, we have applied $\tau_{Nov} = 1.5$, *i.e.,* if σ_{Nov} is 50% larger than R_{Th}, an occlusion is reported through the logic $Novelty$ figure F_{Nov}. It is worth mentioning that through the proposed approach, change of motion direction is reflected in the motion models and hence differentiated from novelty situations.

State Machine

In the proposed system, a 'State Machine' is designed to estimate state of the target in the course of tracking, as demonstrated in Figure 3.6. 'Moving' and 'Stationary' denote the status of a moving or static target, where it is completely visible and normally tracked. In 'FullOcclusion' the object of interest is temporarily not visible in the scene. If the object does not reappear within a certain period of time, state is set to 'Disappear' and the tracking process is stopped. In detecting partial occlusions (our focus here), three scenarios are possible based on the motion models. The three possible partial occlusions, differentiated by the state machine, are named as 'MovOccMov', 'StatOccMov' and 'MovOccStat'. In the first and second case, the target of interest is moving while partially occluded by another moving or stationary object in the scene. In the third case, the target of interest is stationary while partially occluded by another moving object. The partial occlusion state will be triggered, if any of the three states is fired. The state machine transitions are controlled by the target 'Static' probability P_S and $Novelty$ figure calculated through the equations Eqn. (3.12) and Eqn. (3.16). The proposed $Novelty$

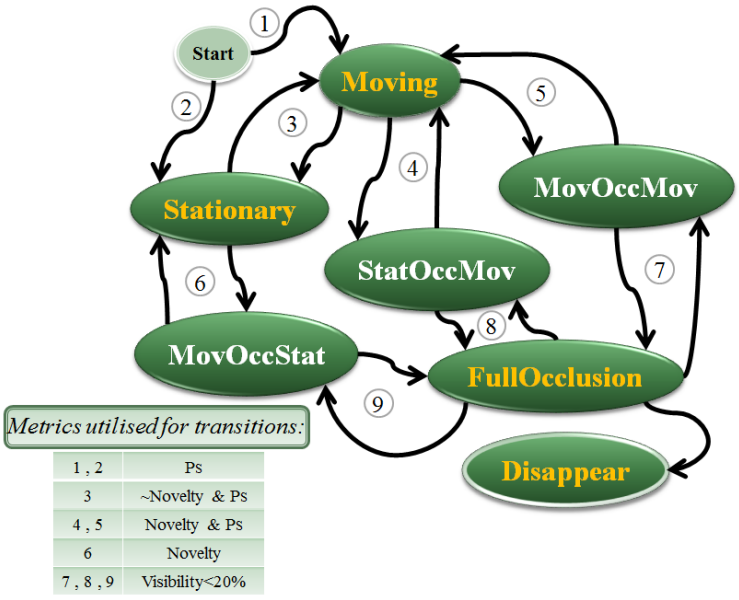


Figure 3.6: Tracker State Machine with transition metrics

figure in Eqn. (3.16) inherently detects partial occlusion situations. While the status probability P_S introduced in Eqn. (3.12), provides a measure of the static energy in the target bounding box. Due to previous target state and the probability P_S, the three partial occlusion states can be discriminated, when $Novelty$ is activated. 'FullOcclusion' may only follow a partial occlusion state. This transition occurs when the target visible area represented by occlusion mask (cf., § 3.4.2) is less than a threshold (in all of our experiments we set the threshold as 20%.).

In the proposed tracker three important mechanisms are controlled according to the target states. Firstly, change of tracking strategy in 'FullOcclusion' state to a short-term tracking by prediction and searching for target energy blob is achieved. This situation frequently happens in surveillance applications, in which a target of interest disappears temporarily. Secondly, an adaptive update strategy for target appearance model is deployed according to the target state. Thirdly, the state machine enables us to accurately capture the dynamics of the target, since only in the 'Moving' and 'Stationary' states (in which the target is fully visible) the Motion Signatures are attained.

Occlusion Mask

In the proposed approach to tracking, the target motion model is utilised to generate an 'Occlusion Mask' which identifies the relevant pixels inside the target window for the Gradient-Descent optimizer during the tracking process. The Occlusion Mask improves the performance of tracking, since the effect of outliers is minimized in the optimization process. This is accomplished by distinguishing and eliminating the irrelevant pixels related to interfering objects in case of occlusion, as well as removing background stationary pixels in the target window. Hence the proposed method is capable of discriminating between objects with different motion dynamics. It also helps to reduce the effect of general aperture problem by removing the pixels in non-textured areas of the target which are devoid of sufficient structure. The aperture problem is about the ambiguity in the motion of non-textured areas, as their motion can not be determined without considering the surrounding structures.

For generating the occlusion mask, the motion signature of each pixel inside the target window is evaluated using the Gaussian functions of the target motion model. Since motion channels with higher average energies represent the status of target motion, we deem to weight them more in generating the occlusion mask. Hence, a linear combination of model functions weighted by target status probabilities (P_S, P_H, P_V) is suggested for computing the occlusion mask as follows:

$$\begin{aligned} M_\omega(\boldsymbol{x}) &= P_S \cdot G_S(\boldsymbol{x}) + P_H \cdot G_H(\boldsymbol{x}) + P_V \cdot G_V(\boldsymbol{x}), \\ G_i(\boldsymbol{x}) &= \mathcal{N}_i(\ E_i(x) \mid \mu_i, \sigma_i\ )\ \ ,\ \ i \in \{S, H, V\}. \end{aligned} \tag{3.17}$$

The problem is that the motion signature of the target pixels provide valid information only in textured areas of the target window. According to equations Eqn. (3.4) and Eqn. (3.5), the oriented energies in non-structured regions drop to zero, while E_O grows. Thus the mask coefficients as calculated by Eqn. (3.17) are invalid in non-texture areas. Hence mask pixels need to satisfy two conditions simultaneously. That is, they should have enough texture and also be relevant to motion model as depicted below:

$$P(x \in M) = P(T_i \cap \bar{O}_i) = P(\bar{O}_i) \cdot P(T_i|\bar{O}_i) \tag{3.18}$$

The probability $P(\bar{O}_i)$ represents the probability that pixel 'x' has enough texture and $P(T_i|\bar{O}_i)$ exhibits the probability that pixel 'x' belongs to target of interest ($x \in T_i$), given it is in a textured area. This probability is M_ω as calculated in Eqn. (3.17). The probability $P(\bar{O}_i)$ may be easily estimated by applying a Sigmoid function to E_O to convert it to probability values. To speed up the computations, we have set a threshold on E_O, to produce a binary mask $M_{\bar{O}}$ for representing pixels with enough texture:

$$\begin{aligned} M_{\bar{O}}(\boldsymbol{x}) = & \quad (E_O(\boldsymbol{x}) < \tau_o) \ , \\ M_R(\boldsymbol{x}) = & \quad (M_\omega(\boldsymbol{x}) > \tau_\omega), \\ M(\boldsymbol{x}) = & \quad M_{\bar{O}}(\boldsymbol{x}) \cdot M_R(\boldsymbol{x}). \end{aligned} \tag{3.19}$$

Hence by utilizing this mask, only the pixels with highly relevant "Motion Signature" (M_R) contribute to the optimization process, while non-textured areas of the target which are prone to the aperture problem will be removed from the mask via $M_{\bar{O}}$. Experiments suggest the parameter values $\tau_\omega = 0.5$ and $\tau_o = 0.05$ yield satisfactory performances.

For illustrative purposes the estimated optimization masks in two different states have been demonstrated in Figure 3.7. As can be seen, the optimization mask removes the majority of irrelevant pixels from the target window through satisfactory discrimination between the object of interest and the undesired interfering objects, as well as realizing the non-textured areas.

3.4.3 Updating The Target Model

In the proposed tracking system, targets are described by an "Appearance Model" and a "Motion Model", which are adaptively updated in the course of tracking. In the following text we elaborate on how this is done for each model.

Figure 3.7: 'Optimization Masks' in two states of Pop-Machines video (York dataset [Can10]) .

Appearance Model

In real world scenarios the appearance of the targets changes over time. This is one of the challenges for tracking algorithms that are dependent on a target appearance model (template) as a matching reference. Hence, a reliable template adaptation mechanism is necessary to maintain an accurate representation of the target. The basic update mechanism applied in this work is an exponential average, which calculates a weighted combination of the previous template $T^{(k)}$ and the optimally aligned candidate $C(x, y)$ in the current frame:

$$T^{(k+1)}(x, y) = \alpha T^{(k)}(x, y) + (1 - \alpha)C(x, y), \quad 0 \leq \alpha \leq 1, \tag{3.20}$$

where the constant α controls the rate of template updating. We note that the appearance model will be updated faster for $\alpha \to 0$ and *vice versa* for $\alpha \to 1$. Unlike conventional blind adaptation schemes, it is possible to employ smaller α here, since template updating is suspended in undesired states. Furthermore, one can adopt different updating rates by utilizing the states seamlessly.

In this work, to prevent corrupting the target template, updating is disabled in occluded states, *i.e.,* α is set to one. In the two states 'Moving' and 'Stationary', we suggest two different updating rates. The template is updated faster in the 'Stationary' state. This is mainly motivated by the obser-

vation that changes in motion direction and pose are more probable through the stationary situation. Based on empirical evaluations, the parameter α is set to 0.9 in 'Stationary' and 0.95 in 'Moving'.

Motion Model

The dynamics of the targets are always changing in the course of tracking as a result of changes in target motion direction and status. Hence the tracking system needs a reliable motion model for the target of interest. For this purpose, similar to the appearance model, an exponential average scheme is utilized. Whenever no novelty is detected in the aligned candidate region, a weighted combination of the previous motion model $\{{\mu_i}^{(k)}, {\sigma_i}^{(k)}\}$ and the current motion model $\{\mu_i, \sigma_i\}$ is assigned as the new motion model of the target in the current frame:

$$\begin{aligned} \{{\mu_i}^{(k+1)}, {\sigma_i}^{(k+1)}\} &= \beta\{{\mu_i}^{(k)}, {\sigma_i}^{(k)}\} + (1-\beta)\{\mu_i, \sigma_i\}, \\ & i \in \{S, H, V\}, \;\; 0 \leq \beta \leq 1. \end{aligned} \tag{3.21}$$

Similarly, the non-rigidity thresholds of the target in three motion channels are updated with a much slower update rate by:

$$\begin{aligned} {R_i}^{(k+1)} &= \gamma {R_i}^{(k)} + (1-\gamma)\sigma_i, \\ & i \in \{S, H, V\}, \;\; 0 \leq \gamma \leq 1, \end{aligned} \tag{3.22}$$

where $\{R_S, R_H, R_V\}$ are non-rigidity thresholds of the target. The non-rigidity thresholds are defined as the expected variance of the motion random variables $\{E_S, E_H, E_V\}$ in the target motion model. Variables $\{{\mu_i}^{(k)}, {\sigma_i}^{(k)}\}$ and ${R_i}^{(k)}$ are initialized by the values of $\{{\mu_i}^{(1)}, {\sigma_i}^{(1)}\}$ in the first video frames since it is presumed that the target starts from a non-occluded state. The parameter values $\beta = 0.7$ and $\gamma = 0.95$ are used for experiments.

3.5 Experiments

In this section, we compare and contrast our tracker against six state-of-the-art trackers in several challenging scenarios. As the focus of this thesis is surveillance applications, the tracker is evaluated on several surveillance datasets with stationary cameras including CAVIAR [CAV04] , PETS2007 [PET07] , i-LIDS [iLI06] and York dataset [YOR10] In the following text, we refer to our proposed tracker as IMSOE, as it works based on 'Intensity' and 'MSOE' features. A video result of the proposed system on i-LIDS scenario is available on 'youtube', for visual demonstration [1].

The reference trackers for our experimental evaluations are the 'SOE Pixel-wise Tracker' [CGW10], IVT [RLLY08], MILTrack [BYB11], Online AdaBoost (OAB) [GGB06], L1-APG [BWLJ12] and L1-IVT [JLY12]. The L1-APG and L1-IVT trackers incorporate an 'Occlusion Detection' mechanism, as elaborated in section § 3.2. To demonstrate the difficulty of selected scenarios, we also assess the performance of the basic mean-shift tracker [CRM03] on gray-scale videos. The size of the targets in our experiments varies in the range of $(25 \times 70 = 1750)$ to $(40 \times 170 = 6800)$ pixels. However based on our experimental observations, the proposed system can perform a reliable tracking for the target sizes above 25×25 pixels. The parameter values used for the experiments are shown in table 3.1.

Table 3.1: Tracking Parameter notations, values and description ;

Parameters	Value	Eq.No.	Description
$\tau_{\mathbf{Nov}}$	1.5	Eq. 3.16	Threshold for Novelty detection
τ_{ω}	0.5	Eq. 3.19	Threshold for motion relevancy
$\tau_{\mathbf{o}}$	0.05	Eq. 3.19	Threshold for Non-textured area detection
$\alpha_{\mathbf{Stat}}$	0.9	Eq. 3.20	Template Update rate in 'Stat' state
$\alpha_{\mathbf{Mov}}$	0.95	Eq. 3.20	Template Update rate in 'Mov' state
β	0.7	Eq. 3.21	Update rate of motion model
γ	0.95	Eq. 3.22	Update rate of non-rigidity thresholds

Since novelty detection in our algorithm plays a significant role for maintaining an up-to-date and valid target template, the precision and recall rates of novelty detection are also presented in the end. To the best of our knowledge, the accuracy of occlusion detection in the recent visual trackers

[1] Occlusion Handling in Single Target Tracking

equipped with this mechanism, hasn't been independently evaluated hitherto. In the literature [MLW+11, KNHH11, WLYY11, BWLJ12], the improvement caused by occlusion detection is demonstrated in terms of the tracking accuracy.

The IMSOE system was initialized using the ground truth box from the first video frame with the assumption that the target of interest in the first frame is non-occluded. Henceforth, frame to frame target motion is captured through parametric motion estimation with respect to intensity template T_i. The Spatiotemporal Oriented Energy filters are implemented via separable convolution. Consequently efficient computation of 'Motion Signatures' is achieved through basic mathematical operations (point-wise addition, squaring and division).

Qualitative Comparison

Five experiments are designed for evaluating the tracking system. In each experiment, our tracker has been compared against six competitors and the mean-shift tracker as a baseline. In the first experiment, the tracker is tested on 'Pop-Machines' video provided in [Can10]. Second and third experiments are designed to test the tracker in two different scenarios from CAVIAR dataset. The fourth experiment is performed on PETS-2007 and the fifth one is carried out with i-LIDS dataset.

Experiment 1 has been conducted on Pop-Machines video which includes two occluding subjects with similar appearance. The man walking from the left is considered as the target of interest, since it is occluded twice during the video sequence in two different states (partial and full occlusion). The subjective results of the SOE, IVT and IMSOE trackers are illustrated in Fig. 3.8 for illustrative purposes.

L1-IVT, L1-APG, IVT, OAB and Mean-shift tracker lose the target after full occlusion. This could be attributed to their tracking strategy as it is not changed in full occlusion (in full occlusion there is no visible target to track). In general, one could expect the aforementioned trackers to find the target after occlusion if

- the duration of full occlusion is short enough

- or a large search area is utilized during tracking.

Needless to say, the former is a limitation and the latter decreases the accuracy of tracker and increases the computational cost drastically. Nevertheless, we will see in the follow-up experiments instances where the competitor trackers are able to find the target after full-occlusion.

Good performance of MIL may be due to capturing multiple instances around the target which models the background around the target to some extent. As shown in Fig. 3.8, SOE tracker demonstrates an unstable situation during the tracking due to the tailing effect, however it doesn't lose the target in the two occlusions. While the SOE tracker is not equipped with an occlusion analyzer, keeping the track of target in full occlusions should be a by-product of the tailing effect. The SOE features are calculated by analyzing a series of previous and future frames. Hence within a couple of frames following the full occlusion event or prior to the reappearance of target in the next frames, its energy blob is observable in the SOE feature space which makes it trackable.

The behavior and performance of IVT, L1-IVT and L1-APG are very similar in this experiment. While these trackers lose the target in the full occlusion, they seem to be smoother and more stable than SOE tracker in partial occlusion and non-occluded situation prior to the full occlusion event. Our proposed tracker shows a stable behavior in this scenario, while following the target in both occlusion events.

Experiments 2 and 3 have been performed on a surveillance video from CAVIAR dataset. The scenario includes two subjects with an occlusion event and severe pose change throughout the video. Some samples of the final tracking results for the SOE, MIL, L1-APG and IMSOE trackers are demonstrated in Fig. 3.9.

In experiment 2 the target of interest is the woman walking toward the camera. The IVT, L1-IVT, L1-APG and IMSOE equivalently perform well among the whole scenario of experiment 2, including the short-term partial occlusion at the end of video. The robustness against the partial occlusion is due to the effect of forgetting factor in updating the IVT appearance model (IVT and L1-IVT) and the occlusion analysis support in L1-IVT, L1-APG and IMSOE trackers.

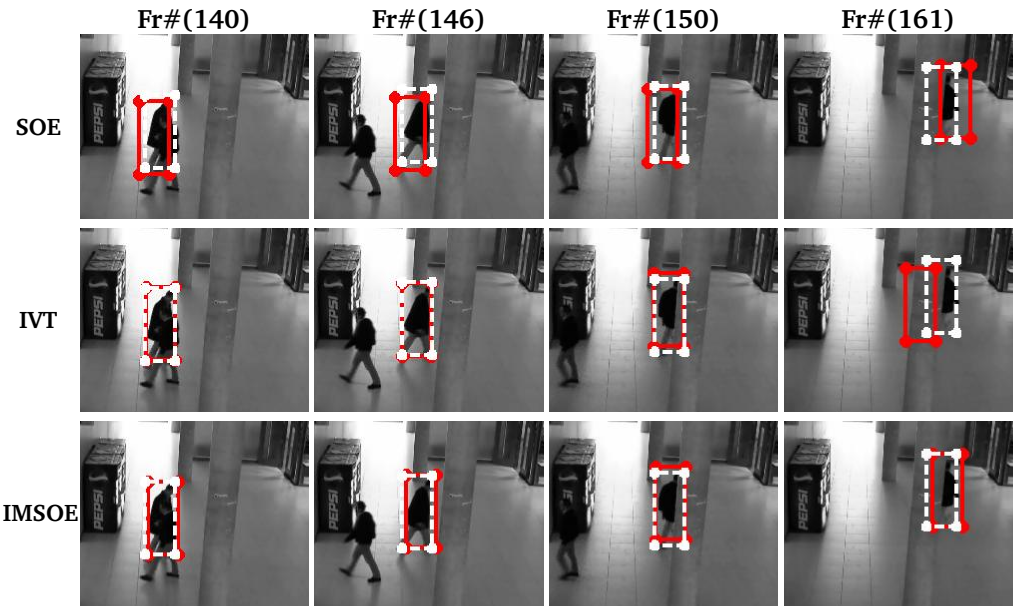


Figure 3.8: Tracking results for Exp.1 (Pop-Machines, York dataset [Can10]). Result of the studied tracker is shown in red and the Ground Truth is shown in white.

The SOE tracker in experiment 2 rapidly loses the target of interest at the final partial occlusion event (Frame#224-Frame#235). Since the woman's location changes very slowly throughout the video, the SOE features of the target window resemble dynamics of a nearly static object. Consequently target dynamics represented by SOE template could be similar to background static areas. This issue causes two main problems for the SOE tracker in this scenario. The first problem is unstable tracking before target occlusion, as the tracker could not sufficiently discriminate between the target and background in some areas. Secondly, in case of occlusion by another moving object, the background areas which have similar SOE features to the template are incorrectly detected as the target of interest. Thus the target will be lost as shown in Fig. 3.9(a).

Experiment 3 is conducted on the walking man in this CAVIAR video. There is no occlusion event in this experiment. However, the target exhibits severe pose variations when the man turns around and moves in the opposite direction. Except for the SOE tracker, Mean-Shift and to some extent the L1-APG, the other five trackers perform roughly well in this scenario.

The L1-APG presents a poor performance in this experiment, probably due to the high similarity of the target-background. Although there is no

occlusion or other special events in the scenario, the L1-APG tracker shows an unstable behavior throughout the video and finally loses the target at the end of experiment. The L1-IVT demonstrates a robust performance which is an indication of its superiority to L1-APG in modeling the target appearance structure. Furthermore, the IVT and L1-IVT trackers perform slightly better than IMSOE in this scenario, especially during the pose and direction variations. This might be attributed to more involved appearance modeling in IVT compared to our simple appearance model. Nevertheless IVT, L1-IVT and IMSOE robustly follow the target of interest through the whole video.

As shown in Fig. 3.9(b), the SOE tracker starts losing the target after the target turns around in the opposite direction. This is mainly because of the rapid changes of dynamic in the SOE features which are used as target template. The general template updating scheme used in the SOE tracker is too slow to follow the drastic template evolutions. This updating scheme is deliberately designed for slow template updating in order to overcome short term occlusions, as the SOE tracker is not capable of discriminating occlusions from normal changes of motion dynamics.

Experiment 4 evaluates performance of the trackers on a video from the PETS-2007 dataset. The scenario is similar to experiment 2, where the target of interest is fully occluded twice. The tracking results for the SOE, MIL, OAB and IMSOE trackers are demonstrated in Fig. 3.10.

Similar to the experiment 2 the IVT, L1-IVT, L1-APG and IMSOE trackers demonstrate a good performance in this scenario. However, IMSOE performs slightly better than the other mentioned competitors. The satisfactory performance for L1-IVT, L1-APG and IMSOE is not surprising, as each of them is armed with an occlusion analyzer. The IVT tracker also performs well due to the short period of occlusion situation and also the effect of forgetting factor in IVT updating mechanism.

On the contrary, the OAB tracker loses the target at the first occlusion event. The MIL tracker follows the target till the second occlusion and gets distracted and consequently loses the target afterwards. As discussed before, learning multiple instances around the target for appearance modeling, helps the MIL tracker to partially model the target background and cope with occlusions. The SOE tracker is easily distracted when the static target of interest

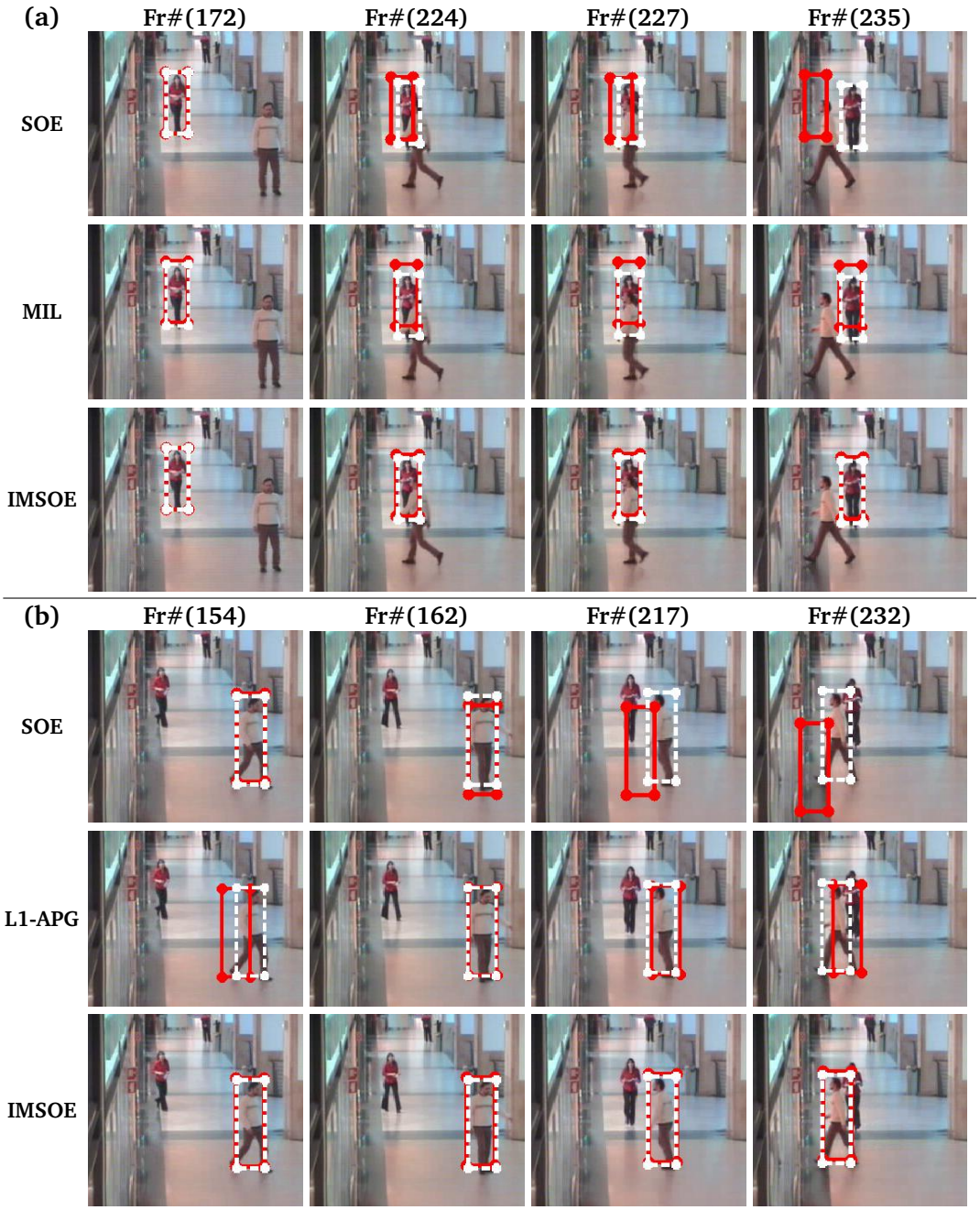


Figure 3.9: Tracking results of (a) Exp.2, (b) Exp.3 for CAVIAR dataset. Result of the studied tracker is shown in red and the Ground Truth is shown in white.

is occluded by the first moving person. This is due to the similarity of the target's dynamics with respect to dynamics of the surrounding static areas.

The appearance of the occluding targets and the target of interest in experiment 4 are distinct. For the L1-IVT and L1-APG trackers, which analyze

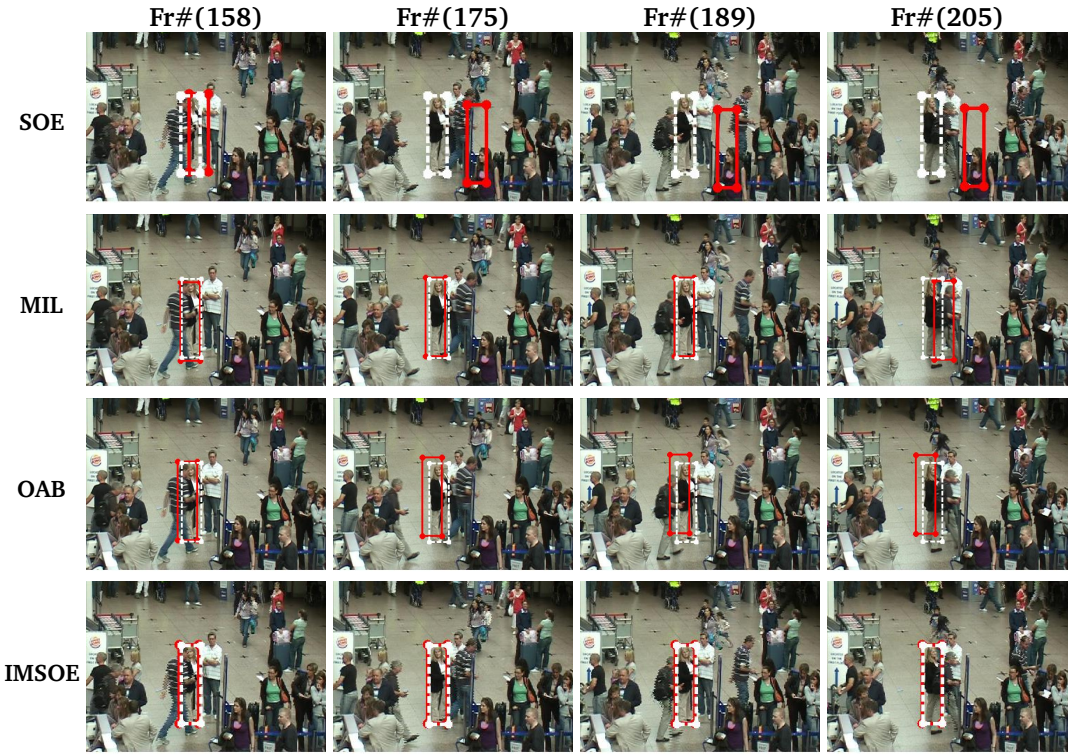


Figure 3.10: Tracking results of Exp.4 (PETS2007). Result of the studied tracker is shown in red and the Ground Truth is shown in white.

the occlusion based on the appearance, we expect to see good performances in this scenario. Nevertheless, L1-IVT and L1-APG don't change the tracking strategy in full occlusion event and will probably fail in full occlusions within a longer period. In the next experiment we consider a scenario in which the occluding objects have similar appearance to the target of interest within a crowded scene and longer occlusions happen.

Experiment 5 is performed on a video sequence from the i-LIDS dataset. The quality of the selected sequence is lower than other videos used in previous experiments. Low illumination and poor quality cause higher similarities in the appearance of the blurry targets. In this scenario several occlusions happen in a crowded scene throughout the video. The tracking results for the SOE, IVT, MIL, OAB, L1-APG, L1-IVT and the proposed IMSOE are shown in Fig. 3.11.

L1-APG and SOE tracker lose the target at early frames of the sequence. However OAB, MIL, IVT and L1-IVT last longer up to the middle of the se-

quence prior to distraction. The SOE tracker is distracted at initial frames of the video sequence due to the similarities between the dynamics of the persons walking together. These two targets cannot be discriminated by the SOE tracker as shown in Fig. 3.11. We note that in the follow-up frames, occluded targets with different motion directions are discriminated by the SOE tracker. The L1-APG also performs poorly and fails to track the target due to the high resemblance of the moving objects' appearance and probably low contrast of the whole scene.

As demonstrated in Fig. 3.11, the IVT and L1-IVT trackers are confused by occlusions in the crowded area. Although the L1-IVT is equipped with an occlusion analyzer, it fails to discriminate among the similar targets of the scene. Furthermore, the forgetting factor of the IVT does not help in this situation, due to the frequent occurrence of lengthy occlusion events. Consequently the IVT based trackers permanently lose track of the target at this point, due to contamination of the template batch and subspace model caused by the occluded frames.

Our proposed tracking system outperforms all the competitors in this scenario, due to the strength of the occlusion analyzer and novelty detector, which discriminate among similar occluding targets and help protecting the target model throughout the video.

Quantitative Comparison

The quantitative performance of the proposed tracking system is compared against seven competitors in this section. Due to low performance of the Mean-Shift tracker in our experiments, we didn't include its qualitative tracking results in previous section. However, the quantitative results are presented in this section as a base-line for comparison.

For objective evaluations, we have utilized center location error, *i.e.,* 'Euclidean distance' between the target and the ground truth centers in the form of precision plots. Beside location error, we assess the performance of all trackers in terms of the MOTP metric [BS08]. Practically speaking beyond a certain boundary around the target, tracking error in terms of distance and position is not sensible, as the target is lost in such situation. Hence the system is actually tracking something else. The MOTP metric measures the

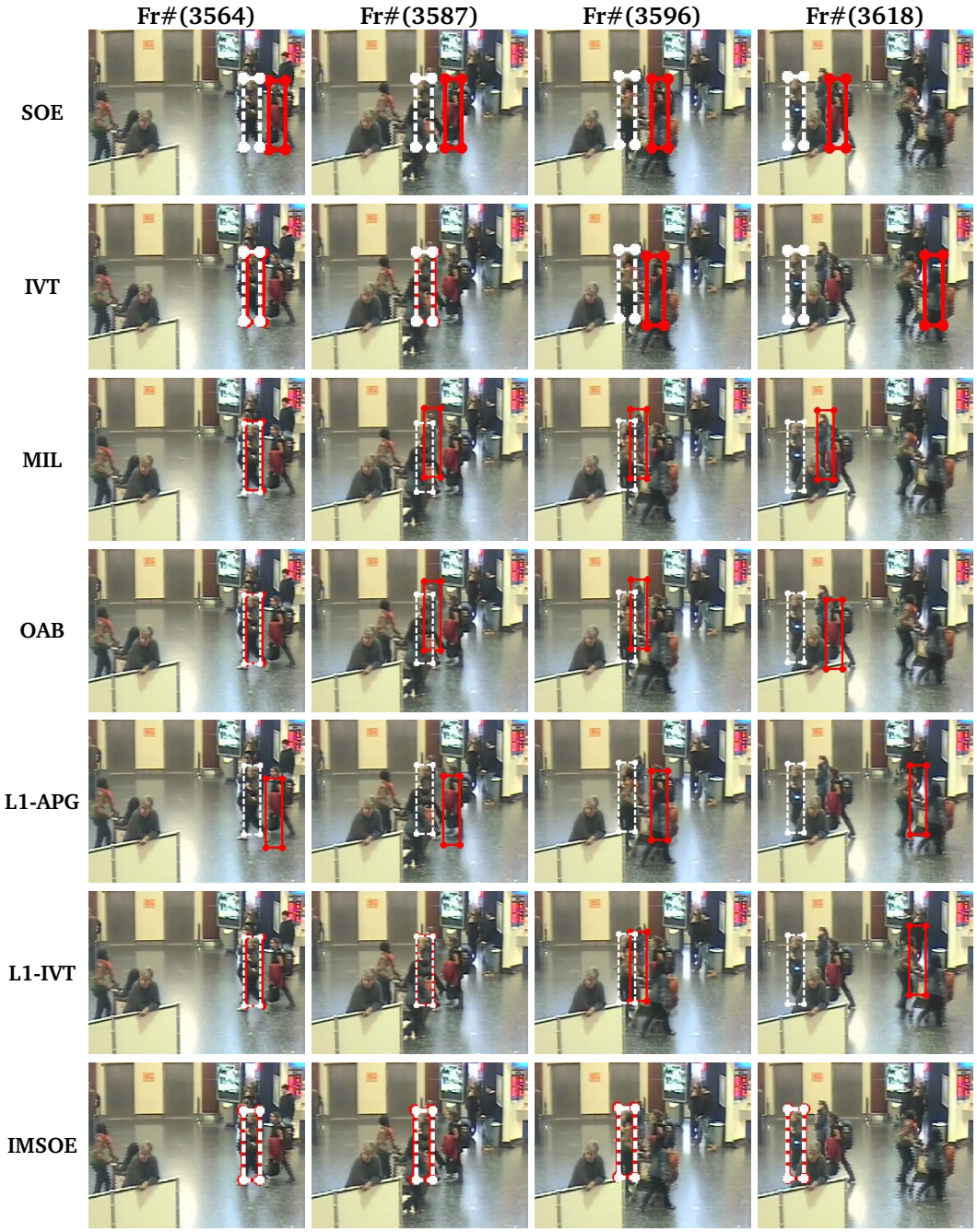


Figure 3.11: Tracking results for Exp.5 (i-LIDS dataset). Result of the studied tracker is shown in red and the Ground Truth is shown in white.

tracking precision by calculating the overlap of the bounding boxes, *i.e.,* the intersection of the estimated box with the ground truth area (see [BS08] for details).

The average center location errors for all the studied methods are shown in table 3.2. We note that the proposed method achieves the lowest error in 3 out of 5 sequences. IMSOE is the second best method for the Exp.2 and the third best in Exp.3, while its performance is marginally close to the winners. The results of Exp.5 demonstrate a large difference between the tracking performance of IMSOE and the competitors, as a result of the target loss in occluded frames of this scenario.

Table 3.2: Average Tracking Error for the studied methods in terms of center location error; Table is color coded for the 1st (green), 2nd (magenta) and 3rd (blue) best results.

	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5
MnShft [CRM03]	33.7	8.0	83.0	13.4	201
OAB [GGB06]	29.3	11.5	6.4	17.6	45.5
MIL [BYB11]	**6.8**	6.4	6.5	17.4	43.2
SOE [CGW10]	**9.1**	5.6	14.4	59.1	63.4
IVT [RLLY08]	17.4	**1.8**	**3.2**	3.3	148
L1APG [BWLJ12]	23.6	**3.4**	5.9	**2.8**	145
L1-IVT [JLY12]	21.6	3.8	**3.0**	**2.8**	134
IMSOE	**5.7**	**2.6**	**3.8**	**2.3**	**2.5**

Table 3.3 shows the average Overlap of bounding boxes in terms of MOTP metric. Based on the MOTP criterion, our proposed method achieves the highest tracking precision in Exp.5, standing significantly above the competitors. IMSOE demonstrates the second best precision in 3 other experiments and stands in the third place in Exp.3. Ranking-wise, it is clear that the proposed IMSOE tracker is the first choice based on both center location and MOTP metric.

Further to the reported average errors, frame by frame tracking errors in terms of location are compared in Fig. 3.12 for all the eight trackers per experiment. As shown in Fig. 3.12, the Mean-Shift tracker failed in the early frames of scenarios 3 and 5 as a result of the low contrast between target and background and also simple appearance modeling. It also performs poorly in scenarios 1, 2 and 4 and fails during the occlusions.

The L1-APG and L1-IVT both incorporate an occlusion analysis scheme and demonstrate decent performances in experiments 1, 2 and 4. However, in experiments 3 and 5, the L1-APG doesn't perform as good as L1-IVT. This

Table 3.3: Average MOTP (Overlap of Bounding Boxes) for the studied methods over all experiments. Table is color coded for the 1st (green), 2nd (magenta) and 3rd (blue) best results

	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5
MnShft [CRM03]	0.42	0.80	0.11	0.80	0.04
OAB [GGB06]	0.47	0.81	0.84	0.68	0.36
MIL [BYB11]	**0.81**	0.84	0.86	0.69	0.36
SOE [CGW10]	**0.65**	0.82	0.64	0.32	0.08
IVT [RLLY08]	0.53	**0.94**	**0.92**	0.90	0.40
L1APG [BWLJ12]	0.53	**0.90**	0.84	**0.95**	0.08
L1-IVT [JLY12]	0.56	0.85	**0.94**	**0.93**	0.44
IMSOE	**0.78**	**0.91**	**0.88**	**0.94**	**0.93**

could be a result of appearance modeling by subspaces in L1-IVT as compared to the pure L1 sparse modeling of appearances in L1-APG. We also note that in the experiments conducted here, L1-IVT performs marginally better than IVT. The advantage of L1-IVT over IVT becomes more significant where long occlusion of targets with sufficiently distinctive appearances occur [JLY12]. In such scenarios even the forgetting factor in IVT updating mechanism, doesn't suffice to protect the subspace model. More specifically, when the occlusion lasts for a considerable amount of time, the IVT algorithm will end up corrupting the batch of target templates as a result of the incremental updating scheme and can never recover from this. The occlusion analysis of the L1-IVT based on sparse coefficients may improve this problem to some extent. Hence incorporating sparse analysis into template updating mechanism of L1-IVT, provides a higher level of protection for the template batch against corruption. However, both IVT and L1-IVT trackers fail in Exp.5, perhaps due to low quality of the video and similar appearance of the occluding objects.

Performance of Novelty Detection

Template contamination is a general problem for all tracking algorithms. In our proposed tracking system such situations are recognized as novelties to prevent corrupting the template with wrong information. Due to the importance of this mechanism, the performance of the 'Novelty Detection System' is separately evaluated in this work. To this end, four sequences used in

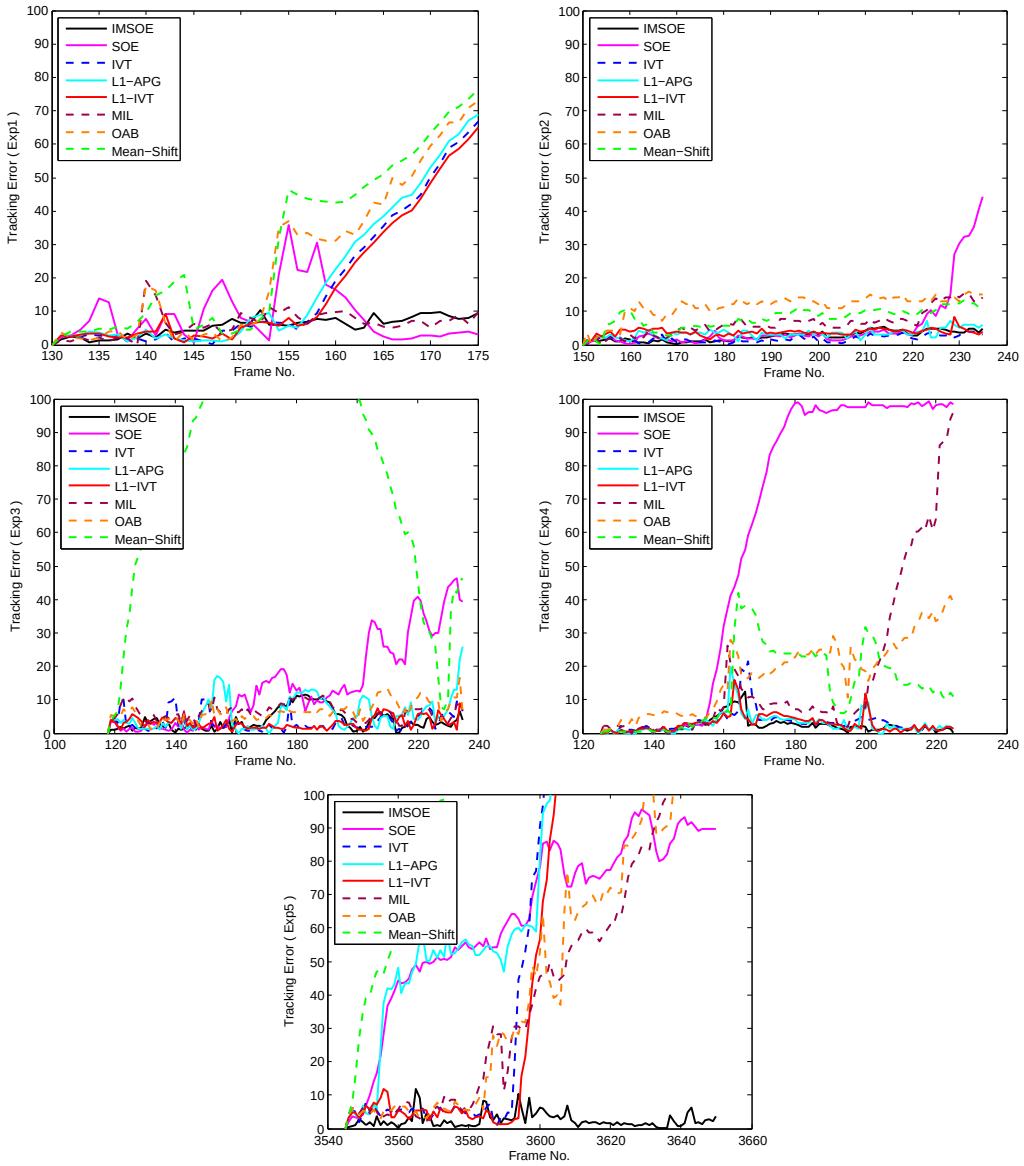


Figure 3.12: Tracking Error diagrams per experiment; Error is the center to center Euclidean distance of the groundtruth and the tracked box

experiments were hand-labeled to provide the ground truth data for target state in every frame. The $Precision = \frac{t_p}{(t_p+f_p)}$ and $Recall = \frac{t_p}{(t_p+f_n)}$ figures for novelty detection were evaluated per experiment, where t_p represents total number of true novelty detections, f_p shows false novelty detections and f_n implies false disregarded novelties. All the target states except 'Moving' and 'Stationary' are considered as 'Novelty' situations, in which updating the tar-

get model is stopped due to partial or full occlusion status. The measured Precision-Recall rates for all experiments are reported in Fig. 3.13, except for 'Exp.3' whose scenario includes no novelty situations.

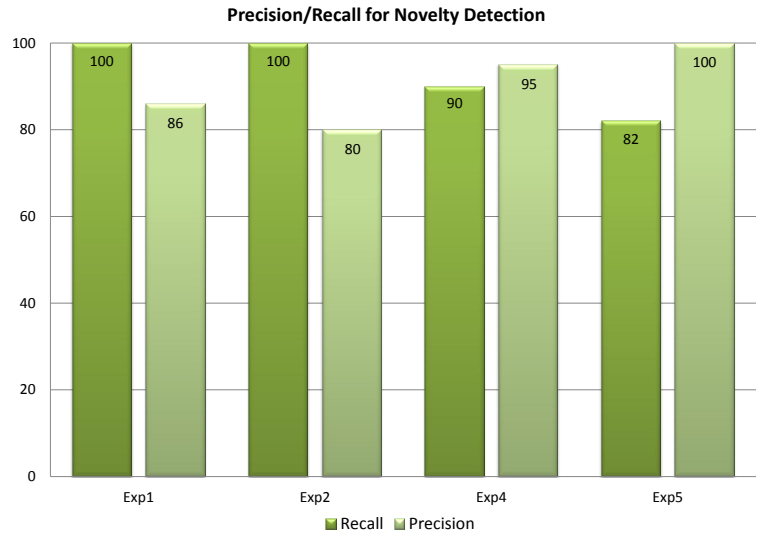


Figure 3.13: Precision/Recall rates of 'Novelty Detection' for experiments Exp1, Exp2, Exp4 and Exp5

Fig. 3.13 reveals that the proposed method detects the novelties with a very high precision. Due to the high performance of the system, only highly reliable information is utilized for updating the target models. Consequently the models are protected against corruptive changes in occlusion situations. The proposed mechanism can offer a great improvement to any tracking system by maintaining a valid target model throughout the process. The tracking results in all of the test scenarios and experiments, illustrate the efficacy of the proposed tracking system.

Novelty detection combined with other trackers

In this part, we show the improvement of another tracking system by the proposed occlusion detection module. Previously, we have demonstrated that the proposed "Novelty Detection" system, enables the base pixel tracker to successfully track non-rigid objects in challenging real world scenarios. To provide further evidence on the benefits of the proposed method, we incorporated the novelty detection component into the state-of-the-art L1-IVT

tracker [JLY12]. For this purpose, the SOE features of the target of interest are calculated in parallel with the L1-IVT tracking process. The extracted energies are utilized by the Novelty Detection modules to perform occlusion analysis and determine the state of the target in the course of tracking.

Upon detecting a 'Full Occlusion', the normal tracking process is paused to avoid distractions. In this situation the target is temporarily tracked by prediction based on the most recent movements of the target. Furthermore updating the target model will be stopped in partial and full occlusion states. Hence, the incorporated modules protect the template batch against corruption in occlusion events and assist the tracker when the target is invisible.

The 'Novelty equipped L1-IVT' tracker was evaluated on our most challenging experiment, *i.e.,* Exp.5, as well as the Exp.1 where the original L1-IVT tracker fails as a result of the full occlusion event. The gain obtained by the novelty detection system for the experiments 2, 3 and 4 is not significant, due to performance of the L1-IVT tracker in these scenarios. The tracking error diagrams of the figure 3.14, demonstrate that the original L1-IVT tracker loses the target of interest in the middle sequences of the video in Exp.1 and Exp.5. However the novelty detection system enables it to perform a robust tracking in the severe occlusions of the scenarios. The average center location errors in Exp.1 and Exp.5 are also reported in table 3.4.

Table 3.4: Average Center Location Errors of ('L1-IVT' *vs*. 'L1-IVT + Nov') in experiments Exp.1 and Exp.5

	Exp.1	Exp.5
L1-IVT	21.6	134.4
L1-IVT+Nov	5.4	3.8

On a related note, our application (and hence the experiments performed in this work) requires the tracking system to deal with non-rigid objects. Since the proposed framework is able to learn the non-rigidity level of the targets, one could expect to gain superior performances when the target is rigid. To investigate this point, we conducted an experiment on the 'Tiger2' sequence [2] which comprised a rigid object. The original L1-IVT tracker failed after 75 frames in this experiment, due to corruption of the target model

[2] `http://vision.ucsd.edu/~bbabenko/project_miltrack.html`

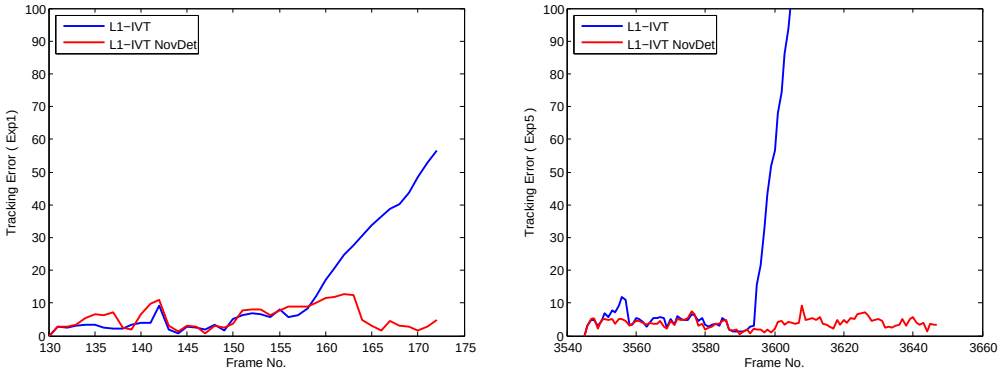


Figure 3.14: Tracking Error diagrams of 'Novelty equipped L1-IVT' *vs*. original 'L1-IVT' for Exp.1 and Exp.5

in consecutive occluded frames. However, a noticeable improvement was observed when the tracker was equipped with the proposed novelty detection component and the system tracked the target successfully.

Tuning the Parameters.

Seven parameters, depicted in Table. 3.1, control the behavior of the proposed framework. We kept the values of these parameters fixed in all experiments performed for this thesis. Nevertheless, we study the effect of each parameter on the performance of the tracking system below and provide heuristic guidelines on how to tune them if it is required for a specific application.

The template update rates, *i.e.,* α_{Mov} and α_{Stat} smooth out stochastic vibrations of the target box, through computing an exponential average of the recent best candidates. Following previous work (*e.g.,* Cannons *et al.* [CGW10]), a value close to one is deemed for the update rates in 'Mov' and 'Stat' states. This results in having a more robust template adaptation, while ensuring that the target representation remains up-to-date, *i.e.,* follows the changes in target appearance. With smaller values of α, any undesired temporary changes will be rapidly introduced into the target template and may cause distraction. For the update rate of Non-rigidity thresholds, *i.e.,* γ, a value close to one is chosen, as the non-rigidity of a target is not expected to evolve much during a scenario. On the contrary, as the target motion dynam-

ics may change rapidly within a few frames, the corresponding parameter, *i.e.,* β, should allow fast updates. In our experiments, we have never observed that choosing a value rather than the suggested ones for (α_{Mov}, α_{Stat}, β, γ) result in significant improvement on the overall performance of the tracking system and one could safely keep the updating values fixed according to Table. 3.1.

As for the threshold values, *i.e.,* τ_{Nov}, τ_{ω} and τ_{o}, the proposed normalization process provides heuristics for success range of each threshold. More specifically, automatic learning of the targets' normal non-rigidity level (Eqs. 3.15 and 3.22), normalized novelty detection metric (Eq. 3.16) and normalization of the SOE features (Eq. 3.6) prior to applying the thresholds, makes the tuning process straightforward. Table 3.5 demonstrates the range of threshold values which were tested for tuning the system. The success range represents the range of parameters for which the tracking system didn't lose the target in our tests. The failure column shows the threshold values for which the tracking has failed. In the success range, we have conducted several experiments for validating the threshold values τ_{Nov}, τ_{ω} and τ_{o}. Each of the thresholds depicted in Table. 3.1 are set as the mid-value of the success range for final assessments. It is worth mentioning that slight changes of the parameters, do not largely affect the performance.

Table 3.5: Parameter Tuning: Range of success and failures ;

	Success Range	Failure Range
$\tau_{\mathbf{Nov}}$	$[1.4 \;\; 1.6]$	$(0, 1.2]\ Or\ [1.8, \infty)$
τ_{ω}	$[0.4 \;\; 0.6]$	$(0, 0.2]\ Or\ [0.8, 1)$
$\tau_{\mathbf{o}}$	$[0.03 \;\; 0.06]$	$(0, 0.02]\ Or\ [0.08, 1)$

Computational complexity

For calculating the target SOE features based on Gaussian Derivatives (*cf.* § 3.3), we need to calculate the three dimensional G_2 and H_2 basis functions across the target. We note that the SOE features are calculated over the target box of size (m, n), not the whole video frame.

The basis functions of 3D Gaussian Derivatives are separable in x, y and t. So we need to apply three one dimensional filters of size p for each basis

function. Please note that only p consecutive video frames are processed together along the time axis t (in this work $p = 9$). Consequently, with six G_2 basis functions and ten H_2 basis functions, $(6+10) \times (3 \times m \times n \times p) = 48mnp$ FLOPs are required to perform the convolutions. A final low pass filter (of length 'p' or less) is also applied to the calculated energy planes (5 channels), to smooth out the energy values across the target. This incurs an additional complexity of $5 \times m \times n \times p$, which brings it up to a total of $53mnp$ FLOPs for computing SOE features. The complexity of Bayesian modeling is negligible compared to computational load of SOE filters. The major calculation in this part is the computation of 'Mean' and 'Standard Deviation' (μ_i, σ_i) for the Gaussian motion models of the target $(\mathcal{N}_S\ ;\ \mathcal{N}_H\ ;\ \mathcal{N}_V)$.

A Matlab implementation of the proposed algorithm can process 10 frames per second on a 3.00GHz Intel machine with a target of size 30×126. However this target size only requires $53mnp = 1.8$ (MFLOPs) for the calculations. The low reported frame rate is due to the Matlab overhead. Hence an efficient C/C++ implementation can easily perform in real time in the case of single target tracking. In order to apply the method to a multiple-target tracking system, the SOE energy features of the whole frame may be required. For a VGA size frame (640×480), the complexity order of the SOE calculation is about 150 MFLOPs per frame, which implies 4.5 (GFLOPS) processing power in a rate of 30 fps. Hence considering the additional required optimizations for tracking or other applications, hardware acceleration will benefit the system for real-time performance.

Table 3.6 provides the performance speed of the studied competitors according to their original papers. For a fair comparison among the trackers, comparable implementations (by Matlab or C++ in similar platforms), should be evaluated on a unique hardware. However, we don't have all the required source codes of the competing trackers or their comparable complexity reports. Some of the available codes are in C/C++ or just executable files (MIL & OAB) and others are in Matlab incorporating Mex functions (IVT, L1-IVT & L1-APG). In spite of these facts, the available information are gathered in Table 3.6 as an overview.

Table 3.6: Computational Complexity of the competitor trackers

	Frame Rate	Year	Notes
OAB [GGB06]	20 (fps)	2006	1.6 GHz CPU, 512 MB RAM
MIL [BYB11]	25 (fps)	2011	C++
SOE [CGW10]	$N.A.$	2010	
IVT [RLLY08]	7.5 (fps)	2008	MATLAB with MEX, 2.8 GHz CPU
L1APG [BWLJ12]	26 (fps)	2012	MATLAB, 3.4 GHz i7 Core CPU
L1-IVT [JLY12]	1.5 (fps)	2012	MATLAB, 2.7 GHz Dual Core CPU, 2GB RAM

3.6 Discussion and summary

In this chapter we introduced a novel approach to take advantage of 'Spatiotemporal Oriented Energy' features for the purpose of robust template tracking in video surveillance applications. The proposed occlusion analysis framework with its three modules ('Novelty Detection', 'Occlusion Mask' and 'State Machine'), provides enough strength for the tracking system to compete with state-of-the-art algorithms. Novelty detection in our system largely improves the template update mechanism and helps to maintain a valid up-to-date target model.

All the experiments in this work are based on the stationary camera assumption. However we believe the proposed system is potentially applicable to moving camera situations, as long as the camera motion does not cause abrupt relocation of the tracked object in the video frames. In the other word, sudden camera motions do not provide suitable grounds to describe erratically moving objects by motion signatures, as no specific motion direction can be defined for objects in such videos.

One important limitation of the current approach is the lack of explicit appearance modeling for novelty detection, as the proposed framework is solely based on motion models. Hence for objects moving roughly in the same direction, the proposed system is not able to detect a novelty situation and the 'Occlusion Mask' generator can not discriminate between the occluding targets. This weak point of the algorithm may cause template contamination and distraction in such scenarios.

Future works: In order to overcome the weak points of the algorithm, we propose to include an efficient appearance model in the framework for the purpose of novelty detection and occlusion analysis. The developed tools may

be applied to modern tracking systems and particle filter based frameworks. Furthermore the framework may be extended to multiple target tracking in template based trackers. More advanced occlusion models may be established for discriminating the occluding targets based on the interaction of the target tracks and the scene geometry. A glimpse of such approach is presented in the next chapter. Re-identification of the lost targets based on effective appearance models is another item to be investigated. This may be applied for recovering the lost targets after long-term occlusions.

Chapter 4

Occlusion Handling in Pedestrian Detection and Tracking

4.1 Introduction

Effective pedestrian detection and tracking under short-term and long-term occlusions is still a challenging research problem in computer vision with many commercial applications. Although state-of-the-art human detectors (e.g., ACF detector [DABP14] and Deformable Part Models-DPM [FGMR10]) have shown promising results in challenging sequences, it is well known that they fail to robustly detect people in the presence of partial occlusions and perform poorly at low resolutions [DWSP09, DWSP12]. Low detection rates for significantly occluded targets, can be considered as an obvious consequence of two technical issues:

- The non-maximal suppression (NMS) procedure for detectors tends to ignore spatially nearby detections (*c.f*. figure 4.1-(e,f)).
- The full body models are inappropriate for detecting partially visible pedestrians(*c.f*. figure 4.1-(d)). .

The above-mentioned issues can be observed in Figure 4.1, which demonstrates the detection results of a state-of-the-art pedestrian detector [DABP14] in various occlusion situations of a PETS-2009 scenario. We note in figure 4.1-(a,d), there is no reported detection for the partially occluded pedestrian in the center of the frame (partial occlusion caused by a fixed scene

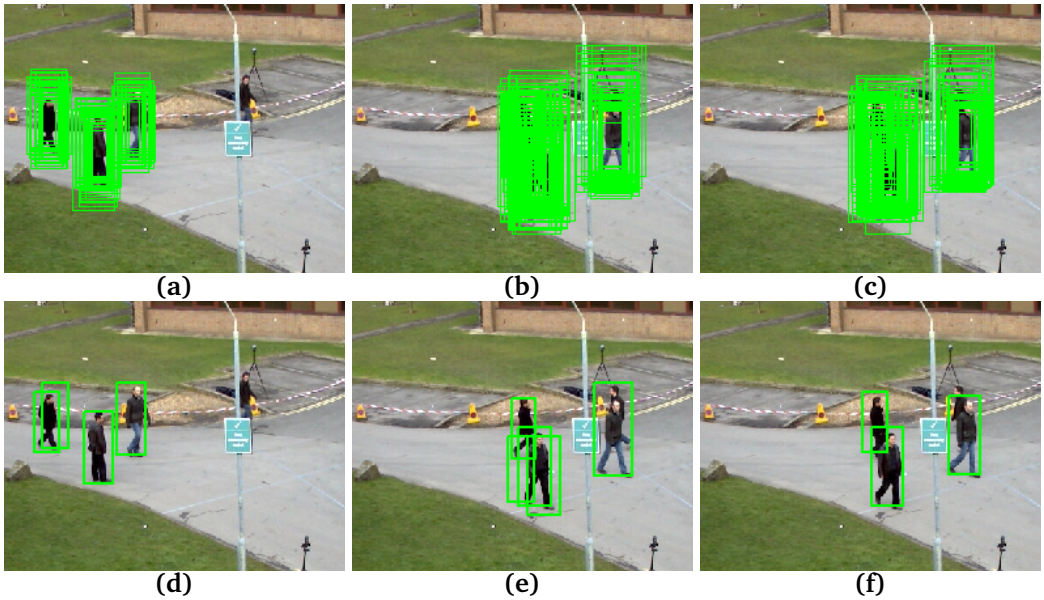


Figure 4.1: (a), (b), (c) All the ACF detections prior to NMS at frames #21, #46 & #47. (d), (e), (f) The ACF detections after NMS (PETS2009, S2L1, V01)

occluder, the lamp post). Furthermore the close detections in figure 4.1-(b,c) are merged to one detection by NMS.

It is generally expected that human detectors which work based on models of body parts (such as the state-of-the-art DPM detector [FGMR10]), be more robust to partial occlussions. However it has been shown that DPM starts to fail at about 20% occlusion for the tested scenarios in [TAS12], while for occlusions beyond 40% true detections become a mere chance.

The problem of tracking by detection for multiple targets has been investigated to some extent in order to improve both tracking and detection accuracy, by taking advantage of the strengths of both [ARS08]. However the task of multiple-object tracking under occlusion is still an open research problem with a few recent attempts around it [MRS14, WTSB12].

Having this in mind, we propose an efficient method to improve over the NMS limitations in occlusion situations and compensate for the unavoidable general detection errors (missed/false detections). Furthermore the system performance does not drop in low resolutions, thus making it suitable for the existing infrastructures and low resolution cameras. The proposed system may also be considered as a multiple-pedestrian detection and tracking, designed for video surveillance applications with static cameras. The system

has shown significant improvement over the state-of-the-art pedestrian detectors [DABP14] for pedestrian analysis in sparse video sequences and performs in real-time on a general CPU. This is a very first step towards solving the more challenging problem of real time crowd analysis. In the next section (4.2), we review some of the related studies and proposed approaches for occlusion handling in pedestrian detection and multiple-person tracking. The proposed framework and the system components are discussed in section 4.3. We finish this chapter by providing experimental results on publicly available pedestrian datasets.

4.2 Related Works

Pedestrian detection in the context of video surveillance can be cast as a multiple target tracking problem. A major trend in multiple tracking systems, is to combine a general object detector with a data association unit, which merges the detection results to establish target tracks. Hence occlusion inference in this context may be pursued in two stages: (1) 'Detection' and (2) 'Data Association'. In other words, developing stronger detectors that are more robust to partial occlusions is the main theme in the first approach, while the second method seeks a better data association scheme.

We continue with a short overview on the first category of systems incorporating occlusion handling in the detection stage. One common approach is to train multiple classifiers for various particular occlusion types. More specifically, different detectors are trained for typical occlusion cases, which are supposed to detect specific body parts such as 'Head, Torso, Legs' [EESG10] or 'Right, Left, Bottom & Upper' body parts [WWRS11]. Then all the detectors are evaluated everywhere on the image and the detection results are merged together, to achieve a more robust performance in various situations. An example of such systems is the Franken-Classifiers [MBTG13], which propose an efficient method for training an exhaustive set of occlusion-specific detectors, rather than the typical 3 to 6 classifiers.

One drawback of such systems is the computation cost of applying several detectors on each frame, which limits their practical usage in real-time appli-

cations. Furthermore, it is well known that the smaller an object model is, the lower will be the detection performance and accuracy. For instance, performance of a 'Head & Shoulder' detector is much lower than a 'Full Body' detector, due to less information contained in the former model. In other words, obtaining a properly modeled part detector, requires sufficient resolution to provide a rich enough representation of the parts. However, in many practical video surveillance applications, such resolution might not be available.

Some studies propose explicit occlusion inference within the detection framework, so that the part detectors can be utilised within a smarter scheme for improving the detection performance [WHY09, GPK11, EESG10, WWRS11]. The main rationale behind such proposals is that the occluded target parts decrease the discriminative power of the detector, resulting in an increased miss rate. Hence improved performance is expected by applying the right detectors on visible target areas.

Various approaches have been suggested for constructing a 3D scene model for estimating the depth and visibility of the objects in the scene. Enzweiler *et al.* [EESG10] use dense stereo and optical flow to estimate occlusion boundaries of the objects based on discontinuities in depth and motion. While Wojek *et al.* [WWRS11] suggest much simpler monocular priors and 3d scene geometry such as common ground plane and objects height for inferring the depth. Depths information are then utilized to determine Objects' visibility and occlusion map. Eventually decision of the detector is concentrated on non-occluded body parts, according to their visibility level. More specifically full-body and component detectors are combined in a mixture-of-experts framework, weighted by the expected visibility of parts [WWRS11, EESG10]. Hence experts are the part detectors and expert weights are proportional to the visibility degree of the associated component. Wang *et al.* [WHY09], proposed another approach for estimating the occlusion map on distinct target cells through utilizing a full-body HOG/SVM classifier. The classification scores of HOG blocks are used to infer the visibility within the cells. Then similar to previous methods, the Part detectors ('Upper body' and 'Lower body') are applied on non-occluded areas of the target, in case of ambiguity and partial occlusion. The major improvement over the state-of-the-art in [WHY09]

is due to the utilizing of the joint features HOG-LBP. Independent performance of the occlusion estimation unit is not presented in the paper, except some sample images. However the improvement offered by the occlusion estimation over the original HOG-LBP is minor, according to the performance curves of the paper. Furthermore the part classifiers alone are not robust enough for calculating the final score, due to their low performance [Sal14].

Contrary to previous approaches for explicit occlusion handling, Tang *et al.* [TAS12] suggest to leverage the person/person overlapping patterns as indicative information for improving the detection performance in crowds. Hence a joint detector is trained to detect pair of occluding pedestrians with various levels of occlusions. With this approach, not only several pairs of occluded pedestrians should be trained, but also generalization ability of the method to handle other occlusion cases is questionable, *i.e.,* occlusions with other types of moving objects and scene occluders need to be trained separately.

The second occlusion handling approach in mutli-target tracking systems, addresses the occlusion problem in the data association stage. The output of a general detector prior to NMS, is fed to the system as input information. Then a data association technique is utilized to estimate the targets' locations in consecutive frames.

In [RC13] an objective function incorporating merely spatial information, is proposed as a replacement to NMS for detection of spatially overlapped pedestrians. The proposed spatial cost function is composed of single detection scores and pairwise overlap constraints. The proposed system demonstrates some improvement over simple NMS in detection of overlapping targets, when it comes to deciding which overlapping detections should be suppressed. However the temporal information is completely ignored in this framework and the missed detections/false alarms still remain unsolved.

Despite all progress in pedestrian detection, undesired errors such as missed detections and false alarms are still unavoidable, especially in presence of occlusions. Such ambiguities can be resolved by taking advantage of temporal information in data association stage. Perhaps one of the simplest approaches for feeding the temporal information to the system is proposed by Brown *et al.* [BFP14]. In this work a temporal Non-Max-Suppression

(tNMS) is proposed to compensate for spatial overlaps of the targets. The detection results of n consecutive frames are combined in a batch. Then a standard NMS is applied on the whole batch to produce the final detection results for the middle frame. In spite of the improvement over the standard NMS, false positives problem still remains unsolved.

Other examples among data association methods, propose some cost function to perform spatial and temporal association among the detection results. Spatial association estimates optimal states for the targets based on the distribution of detection results in the frame. Temporal association explores the correspondence of objects across-time and tends to estimate smooth target tracks. For instance, Milan(Andriyenko) *et al.* [AS11, MRS14] propose an energy cost function which incorporates constraint terms on observation/detection evidence, targets' appearance, smooth motion and collision avoidance. The motion term is a constant velocity model which encodes the distance between target velocity in consecutive frames. An occlusion model is also integrated in the observation term of the global objective function, to penalize existing targets with no evidence. However the proposed occlusion model burdens a heavy computation load on the system, cutting down the system performance to one frame per second. The proposed objective function is highly non-convex due to several ad-hoc energy terms involved. Consequently the suggested gradient descent optimization, largely depends on good initialization and improvised sampling heuristics to avoid local minima.

Andriluka *et al.* [ARS08] propose a probabilistic tracking-by-detection framework rather than utilising a global data association cost function. The proposed data association scheme incorporates three stages to exploit temporal coherency in short, middle and long periods. Initially the position, scale and rough articulation of the body parts is estimated with a part-based model in single frames. Then dynamics of the individual limbs are modeled with a hierarchical Gaussian process latent variable model (hGPLVM), to obtain temporal coherency within a walking cycle (tracklets consisting 6 consecutive frames). In the second step, a hidden Markov model (HMM) is utilised to extend the tracklets to longer people tracks through a recursive Viterbi algorithm, between major occlusion events. The suggested HMM

works based on a generative appearance model extracted from tracklets and a dynamical model composed of Gaussian position dynamics and the GPLVM articulation dynamics. In the third step, the generated tracks are associated using the appearance model and a coarse motion model, to track people over even long periods of time. The proposed part-based model and the dynamic model of the limbs over a walking cycle, provide a principled way to handle partial occlusions. However such detailed representation requires sufficient resolution to properly model the parts' appearances. The Computational cost of the complex process and the high resolution requirement of the system, restricts its practical value for real-time applications.

We believe more efficient solutions are required to perform robustly on existing video surveillance infrastructures given their practical limitations, such as limited resolution, limited processing resources along with the real-time speed requirements.

4.3 Technical Approach

We propose a data driven spatio-temporal clustering framework which leverages the consistency in the motion and scale of the tracked targets across frames. The system is computationally very efficient and provides a real-time performance. The proposed cost function helps to maintain a smooth track for every target. It is also able to handle occlusion situations when the targets have different scales or motion directions. The method can be seen as an improved adaptive Non-Max-Suppression (NMS) method, which is aware of the number of existing targets in the scene and provides a solid performance for highly overlapped targets. Moreover, we have taken advantage of the notions 'Depth/Height Map', the 'Scene Entry/Exit' and an 'Overlap Matrix' to post-process the results of the optimization and consolidate the estimated tracks. The track consolidation module takes care of integrating the incomplete tracks which don't start or end at Entry/Exit areas. It also removes spurious tracks with low confidence.

Figure 4.2 demonstrates a high level block diagram of our proposed pedestrian analysis system. The input to system is the detection results of a pedestrian detector prior to NMS, which are processed by our data association

framework. The proposed system optimizes a track for each target within a clustering framework, while it realizes new pedestrians entering the scene. Hence a new cluster is defined for every emergent target within the frames, which is constantly updated along the video sequence. The module dedicated to 'New Cluster Evaluation', applies NMS on non-associated members reported by 'Clustering Unit' and evaluates the NMS results with a confidence score (based on Depth-Height Map & Detection Frequency), in order to instantiate additional clusters for the new entrant targets. The above mentioned clustering cost function along with the proposed scheme for cluster instantiation, seems to be adequate enough to mitigate the need for Foreground/Background modeling towards reducing the false positives. The proposed framework significantly improves the detection performance in sparse crowds, by removing most of the false positive detections.

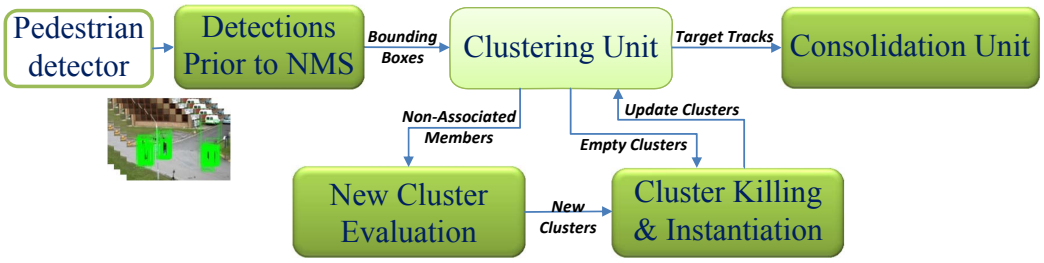


Figure 4.2: High-level Block Diagram of the Multi-Pedestrian Tracking System

4.3.1 Data association framework

We first describe our technical approach on how to merge multiple detections in frame t, given a known number of objects (N) in the frame. The problem is cast as estimating accurate states of the targets through robust clustering of the raw detection results. We define our data driven loss function based on a bounded Euclidean distance measure, which is inspired by the work of Burgos *et al.* [BHPD13] on pose estimation. Let $(X^t, S^t) = \{(x_i^t, s_i^t)|\forall i,\ 1 \leq i \leq n^t\}$, be the n^t detector estimations and their relevant detection scores at frame t. A cluster is defined for every target in the frame and $Y^t = \{y_j^t|1 \leq j \leq N\}$ represent the cluster centers at frame t. Then a data loss term related to the cluster predictions $Y^t = \{y_j^t \in \mathbb{R}^D, 1 \leq j \leq N\}$,

is defined as below:

$$L_{Data}(Y^t) = \frac{1}{S^t} \cdot \sum_{i=1}^{n^t} s_i^t \cdot \min_j d_b(x_i^t, y_j^t) \tag{4.1}$$

$$S^t = \sum_{i=1}^{n^t} s_i^t \ , \ d_b(x, y) = \min(\tau, \|x - y\|_2^2))$$

where d_b is an Euclidean distance bounded at a maximum threshold τ and data loss function L_{Data} is normalised by summation of all the detection scores, S^t. The threshold τ represents the extent (maximum radius) of the identified clusters. The constant threshold τ can be set according to the physical constraints of the environment as the average width of the objects in the scenario. This threshold is comparable to the overlapping threshold in Non-Max Suppression (NMS) post-processing of the object detectors. Suppressing the overlapped targets in occluded scenarios is a major drawback of the NMS process. However we will show that the proposed framework minimizes this undesired drawback of the NMS method, due to the system awareness about the number of existing targets in a local area, which is leveraged to prevent suppression of overlapped targets. Apparently, introducing incorrect number of targets to the system may mislead the optimisation framework. Hence, good performance of the system is dependent on having a reliable estimate about the number of existing targets and their initial states. The proposed method for target detection and new cluster instantiation among the scenario, is discussed in section 4.3.2. This approach leads to improved occlusion handling in certain situations.

In order to minimize the loss function in Equation (4.1), each cluster center is updated to the weighted mean state of its own members at each step. The cluster members are within a distance τ of the cluster center y_j^t based on definition. More specifically, a member $\{x_i\}$ pertains to cluster y_j if it satisfies the following conditions:

$$(x_i \in y_j) \ if: \begin{cases} \|x_i - y_j\|_2 < \|x_i - y_k\|_2 \ , \ \forall k \neq j \\ \|x_i - y_j\|_2 \leq \tau \end{cases} \tag{4.2}$$

The proposed clustering cost function in Equation 4.1, is in spirit similar to

the standard weighted K-Means. However its behavior is practically different, as it performs locally within the bounding threshold τ to minimize the effect of outliers. This can be seen more clearly if we rewrite Equation (4.1) in a linear form by substituting an indicator function ($\mathbb{1}_A$) for the nonlinear '$\min_j$' function:

$$L_{Data}(Y^t) = \frac{S^t_{\bar{a}}}{S^t} + \frac{1}{S^t}\sum_{i=1}^{n^t}\sum_{j=1}^{N} m^t_{ij} s^t_i \, \|x^t_i - y^t_j\|_2^2, \tag{4.3}$$

$$m^t_{ij} = \mathbb{1}_A\big(\|x_i - y_j\|_2 \leq \|x_i - y_k\|_2, \, \forall k \;\; \& \;\; \|x_i - y_j\|_2 < \tau\big),$$

where $S^t_{\bar{a}}$ is sum of the detection scores s^t_i that are not associated with any cluster.

In the standard k-means, each step is optimal due to the well-known mathematical lemma which states: The function $f(y_j) = \sum_i w_i \|x_i - y_j\|_2^2$ is minimized with respect to variable y_j by substituting $y_j = \sum_i w_i x_i / \sum_i w_i$ in the function. However in the proposed bounded K-Means, this replacement is only guaranteed to be locally optimal, *i.e.,* the estimated y^t_j will be the optimal solution within the τ neighborhood of its recent location. Intuitively speaking, $S^t_{\bar{a}}/S^t$ appears as a constant in the equation and does not play a role in the optimization. However with each update of the cluster center, the cluster members and $S^t_{\bar{a}}$ are prone to change, which implies an iterative operation until a steady state is reached. The convergence occurs when the cluster center and thus the cluster members are settled down and don't change further. To escape from the local traps, frequent execution of the algorithm with various different initializations, is a common approach which of-course implies an excessive computation load. However the proposed initialization scheme in our framework (discussed in subsection 4.3.2), has practically demonstrated enough robustness to mitigate the need for random initialization and frequent execution of algorithm.

As mentioned earlier, the proposed clustering framework has another advantage over the standard NMS in inter-person occlusion situations. It is known that standard NMS suppresses the overlapping detections within a fixed surrounding area (*c.f.* figure 4.3, Fr#46-b & Fr#47-b). However as demonstrated in figure 4.3 (rows a & b), the extent of the clusters in the pro-

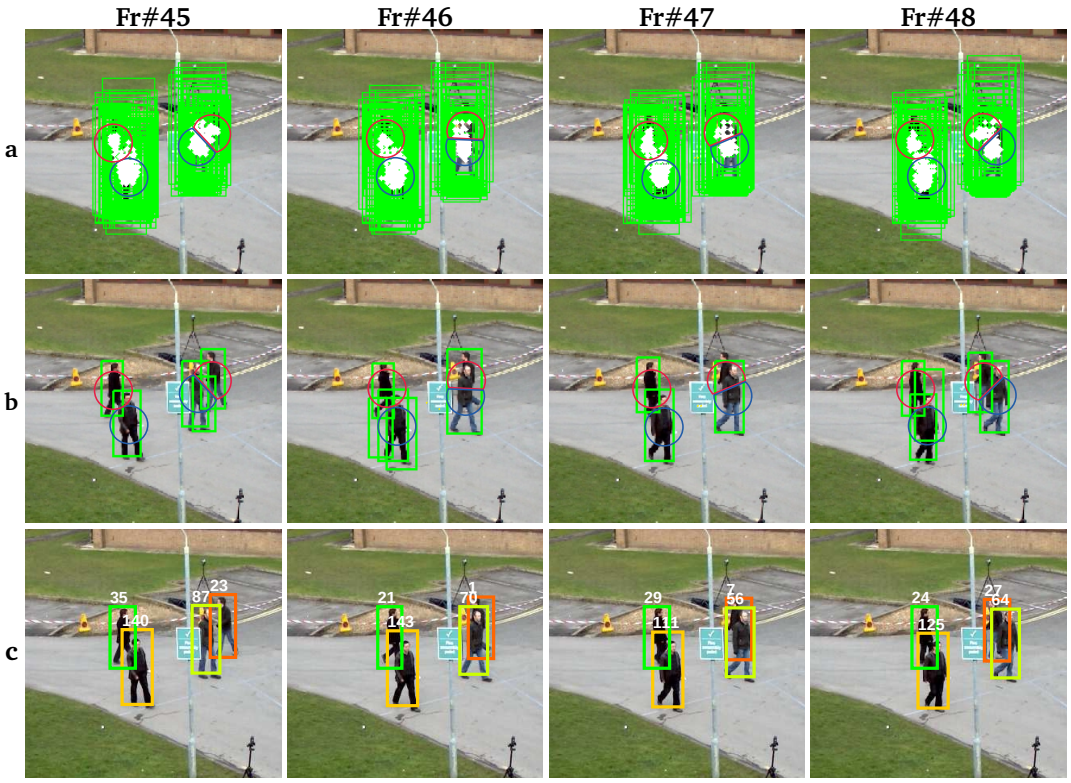


Figure 4.3: (a) & (b): Green bounding boxes demonstrate the ACF detector results before and after NMS; White dots are the bounding box centers; Blue and Red circles show the estimated target clusters by the system; (c): The final results of our proposed clustering framework

posed framework is intrinsically adjusted depending on the proximity of the existing clusters in the scene. This unique property is enabled by the competing term in the definition of membership function m_{ij}^t in Equations (4.3) and (4.2) ($\|x_i - y_j\|_2 < \|x_i - y_k\|_2$). Hence, although the threshold τ is fixed, this term makes the cluster boundaries flexible in occlusion events.

The clusters shown in figure 4.3-(a,b), are a rough projection of the real clusters on (x-y) plane for visualization, as the target clusters and features have four dimensions in our system. The threshold τ determines the maximum extent of the clusters. However the inherent competition among the data members for associating to overlapping clusters (Equation (4.2)), reduces the neighboring threshold among them (*c.f*. figure 4.3, rows a & b). In this case the members in the overlapping area are divided into groups based on their proximity to the cluster centers. In other word, each of the data members in the overlapping area are associated to only one cluster which has

a closer center. This is a result of the system awareness about the number of existing targets in the overlapping area. This approach might lead to inaccurate localization of the clusters in some occlusion situations. However the system maintains the recognized identities rather than killing or suppressing an occluded target cluster. Furthermore, the localization uncertainty of the overlapped targets will be improved by introducing temporal terms into the data association framework, which will be discussed later. Further improvement can be achieved by utilizing appearance models of the existing targets in the clustering algorithm. However this is considered for future extensions of the proposed system.

More interestingly, as demonstrated in figure 4.3, reduction of the cluster radial extent occurs only along the orientation that the two clusters are overlapping, while for other directions the cluster members may spread upto the maximum radial distance τ. In terms of NMS process, we can think of variable directional overlapping thresholds for different situations. This property occurs due to the competing term of the membership function m_{ij}^t in Equation (4.3), as discussed earlier. In an occlusion event, the neighboring threshold of the overlapping clusters are flexibly reduced along their overlapping direction, due to the competition among the data members, induced by membership function m_{ij}^t. Hence upon establishment of a new cluster for an emerging target, the system does not kill the cluster due to short term occlusions or missed detections in the middle of scenario. Even if two clusters are severely occluding each other, non of them will be suppressed. However, there might be some localization inaccuracy in such cases.

In the next step, we need to attain temporal consistency among the subsequent detections throughout the video sequence. To this end, it is common to involve smoothing terms across multiple frames in the loss function. So let's consider two additional terms, namely a 'Constancy' term and a 'Smoothness' term, in our loss function as below:

$$L_{Cnst}(Y^t, Y^{t-1}) = \frac{1}{N}\sum_{j=1}^{N} \|y_j^t - y_j^{t-1}\|_2^2, \tag{4.4}$$

$$L_{Smth}(V^t, V^{t-1}) = \frac{1}{N}\sum_{j=1}^{N} \|v_j^t - v_j^{t-1}\|_2^2,$$

where $v^t = {}^{d_y}/_{d_t} = (y^t - y^{t-1})$. Then the overall loss function over a specific interval becomes:

$$L(Y) = \sum_{t=t_1}^{t_2} \left\{ L_{Data}(Y^t) + \lambda_1 L_{Cnst}(Y^t, Y^{t-1}) + \lambda_2 L_{Smth}(V^t, V^{t-1}) \right\}, \quad (4.5)$$

The constants λ_1 and λ_2 controls the influence of temporal terms. The 'Constancy' term in the loss function tends to keep the estimations in consecutive frames close to each other, encouraging a constant target state. On the other hand the 'Smoothness' term boosts a smooth target motion, encouraging a constant velocity for the targets. Although the proposed terms can both potentially improve the localization performance, they are contrary by definition. Apparently the two terms may cancel out the effect of each other in the loss function, as they encourage constancy and smooth motion simultaneously.

We propose a framework to take advantage of both temporal terms, in a way that their stabilization influence on the estimation process will be accumulated towards handling occlusion situations. More specifically two types of indications are leveraged in our system for occlusion handling:

- While the nearby overlapping detections are consistently merged in the framework, overlaps at far different scales should not be fused.
- The overlapping targets moving in different directions, can be separated due to their different motions.

These indications are formulated in the 'Data' term and 'temporal' terms of the proposed cost function. To provide a more clear intuition about the objective of the proposed loss terms in the cost function (discussed later), we note that:

- The data term in the clustering function, naturally separates very far points in position or scale through the introduced distance metric.
- The 'Constancy' temporal term of the cost function further improves the separability on the scale axis.
- The 'Smoothness' temporal term in the cost function provides the basis

for discriminating the overlapping targets with different motion directions, even in close scales.

In order to achieve the mentioned goals, we propose to split the feature vector into separate parts and apply the temporal terms on relevant chunks. More specifically the feature vector used in detection results $X^t = \{x_i^t \mid 1 \leq i \leq n^t\}$ and cluster states $Y^t = \{y_j^t \mid 1 \leq j \leq N\}$ contains two types of information. Let's define two functions f_P and f_S to retrieve 'Position' and 'Size (Scale)' information from the feature vector: $f_P(X) = (x, y)$ and $f_S(X) = (w, h)$ (or $f_S(X) = s$, representing the scale number in the feature vector). Then $X = (f_P(X), f_S(X))$, $Y = (f_P(Y), f_S(Y))$ and we define the temporal terms for the loss function as below:

$$\mathscr{L}_{Cnst}^{S}(t) = L_{Cnst}(f_S(Y^t), f_S(Y^{t-1})) = \frac{1}{N}\sum_{j=1}^{N} \|f_S(y_j^t) - f_S(y_j^{t-1})\|_2^2 \,, \tag{4.6}$$

$$\mathscr{L}_{Smth}^{P}(t) = L_{Smth}(f_P(V^t), f_P(V^{t-1})) = \frac{1}{N}\sum_{j=1}^{N} \|f_P(v_j^t) - f_P(v_j^{t-1})\|_2^2 \,,$$

where $f_P(v_j^t) = f_P\left(\partial y_j / \partial t\right) \approx f_P(y_j^t) - f_P(y_j^{t-1})$. Hence the 'Constant' term is defined based on the scale/size features and 'Smoothness' term is defined based on spatial position. The introduced temporal terms $\mathscr{L}_{Cnst}^{S}$ and $\mathscr{L}_{Smth}^{P}$ encourage constant size and smooth motion (constant velocity) in adjacent frames respectively and collaboratively improve the robustness of the system. Apparently applying each of the temporal terms on the irrelevant chunk of the feature vector negatively affects the system performance. For the special case of target motions along the camera view line (for example towards the camera), the above mentioned assumptions are still valid, due to the fact that the temporal terms are defined on few adjacent frames. In the light of above, we define our loss function as below:

$$\mathscr{L}(Y) = \sum_{t=t_1}^{t_2} \left\{ L_{Data}(Y^t) + \lambda_1 \mathscr{L}_{Cnst}^{S}(t) + \lambda_2 \mathscr{L}_{Smth}^{P}(t) \right\}, \tag{4.7}$$

The constants λ_1 and λ_2 are experimentally tuned, to control the influence of temporal terms towards an acceptable occlusion handling performance. The empirical values used in our experimental setup are $(\lambda_1, \lambda_1) = (10, 10)$.

However our observations demonstrate that the system is not very sensitive to the exact parameter values, *i.e.,* 25% variation in λ_1 and λ_2 do not cause a sensible change in the system results.

Now we need a solution for minimizing the cost function $\mathscr{L}(Y)$ in Equation (4.7). Due to linearity of the Euclidean distance metric, we know that $\|x-y\|_2^2 = \|f_P(x-y)\|_2^2 + \|f_S(x-y)\|_2^2$. Hence the 'Data' term can be split in two parts. By using Equation (4.3), the loss function is formulated as:

$$\mathscr{L}(Y) = \frac{S^t_{\bar{a}}}{S^t} + \sum_{t=t_1}^{t_2} \{\mathscr{L}^S_{Data}(t) + \lambda_1 \mathscr{L}^S_{Cnst}(t)\} + \sum_{t=t_1}^{t_2} \{\mathscr{L}^P_{Data}(t) + \lambda_2 \mathscr{L}^P_{Smth}(t)\},$$

$$\mathscr{L}^{\phi}_{Data}(t) = \frac{1}{S^t} \sum_{i=1}^{n^t} \sum_{j=1}^{N} m^t_{ij} s^t_i \, \|f_\phi(x^t_i) - f_\phi(y^t_j)\|_2^2 \ , \ \forall \phi \in \{P, S\}, \tag{4.8}$$

We note that the two parts of the feature vectors retrieved by (f_P, f_S) are independent from each other. Hence the two summations in Equation (4.8) are independent, due to independence of their variables. Consequently for minimizing the loss function $(\alpha + \mathscr{L}_1 + \mathscr{L}_2)$ in Equation (4.8), the two terms $\mathscr{L}_1$ and $\mathscr{L}_2$ can be independently optimized. Each part of the target state vector $(f_P(Y), f_S(Y))$, will be estimated by optimizing the relevant terms $(\mathscr{L}_2, \mathscr{L}_1)$. We perform the optimization through a frame-by-frame strategy, *i.e.,* the optimum Y^t on frame t is estimated, while all the other frames are frozen with fixed clusters (Y). To this end, we first rearrange the independent loss functions $(\mathscr{L}_1, \mathscr{L}_2)$ in Equation (4.8) in a form to contain the terms depending on Y^t. Then $\mathscr{L}_1$ and $\mathscr{L}_2$ can be minimized with respect to Y^t.

$$\mathscr{L}_1(Y^t) = \frac{1}{S^t} \sum_{i=1}^{n^t} \sum_{j=1}^{N} m^t_{ij} s^t_i \, \|f_S(x^t_i - y^t_j)\|_2^2 \ + \ \cdots \tag{4.9}$$

$$+ \frac{\lambda_1}{N} \sum_{j=1}^{N} \left\{ \|f_S(y^t_j - y^{t-1}_j)\|_2^2 + \|f_S(y^t_j - y^{t+1}_j)\|_2^2 \right\},$$

$$\mathscr{L}_2(Y^t) = \frac{1}{S^t} \sum_{i=1}^{n^t} \sum_{j=1}^{N} m^t_{ij} s^t_i \, \|f_P(x^t_i - y^t_j)\|_2^2 \ + \ \cdots \tag{4.10}$$

$$+ \frac{\lambda_2}{N} \sum_{j=1}^{N} \left\{ \|f_P(v^t_j - v^{t-1}_j)\|_2^2 + \|f_P(v^t_j - v^{t+1}_j)\|_2^2 + \|f_P(v^{t+1}_j - v^{t+2}_j)\|_2^2 \right\},$$

where $v_j^t = y_j^t - y_j^{t-1}$. Through factorizing $\sum_{j=1}^{N}$, formulation of $\mathscr{L}_1$ and $\mathscr{L}_2$ will turn into a standard K-Means problem:

$$\mathscr{L}_1(Y^t) = \sum_{j=1}^{N}\sum_{i=1}^{n^t+2} \widehat{s_i} \cdot \|f_S(y_j^t) - f_S(\widehat{x_i})\|_2^2 \;\Rightarrow\; f_S(y_j^t) = \frac{\sum_i \widehat{s_i} \cdot f_S(\widehat{x_i})}{\sum_i \widehat{s_i}}$$
$$\mathscr{L}_2(Y^t) = \sum_{j=1}^{N}\sum_{i=1}^{n^t+3} \widehat{s_i} \cdot \|f_P(y_j^t) - f_P(\widehat{x_i})\|_2^2 \;\Rightarrow\; f_P(y_j^t) = \frac{\sum_i \widehat{s_i} \cdot f_P(\widehat{x_i})}{\sum_i \widehat{s_i}} \quad (4.11)$$

The proposed optimization will be executed recursively on each frame up to the convergence point (usually a couple of times suffice). By using a reliable initialization scheme in our system which provides decent estimate of the existing targets in the scene, we don't need to use random initializations and sweep forward and backward among the video frames to reach an optimal solution (as suggested in [BHPD13]). During the optimization, all the clusters/tracks are constantly monitored and the empty clusters that are void of any members for a minimum number of frames (set to 5 in our experiments), are considered as Zombie centers and killed. Concurrently the non-Associated members are monitored throughout the optimization process for instantiating new clusters. As soon as new targets appear in the frame, the initialization unit instantiates new clusters. Hence the number of targets are robustly estimated during the process. Since the temporal terms in the loss function incorporate two neighboring frames, the optimization is sequentially performed on small space-time volumes of the video. The process sweeps the entire video sequence frame by frame, in order to form all the targets' tracks simultaneously. Furthermore one round of K-means optimization on the whole video sequence, provides an initial reliable estimation of the tracks within the scenario.

4.3.2 Cluster instantiation and Track Consolidation

In this section we address two important processes in our multi-pedestrian tracking system. The first critical mechanism is the cluster instantiation problem during optimization, due to the fact that the good performance of the proposed clustering optimization is highly dependent on a reliable initial states. The second important process, is the final stage of 'Track Consoli-

dation'. Following the extraction of the target tracks in the video, a post processing is required to consolidate the incomplete tracks, remove the spurious tracks and refine the final solution. Details of the proposed initialization scheme and the track consolidation framework, are described in this section.

Cluster instantiation

Without a reliable initial estimation of the target states, we need to run the system with various random initializations, to increase the chance of finding an optimal solution. Apparently this approach reduces the certainty of the system, while slowing down the process due to the extra computational load. In order to obtain a reliable initial state for our optimization framework, we propose to use a Non-Max-Suppression (NMS) approach combined with a composite confidence metric. The proposed confidence metric largely reduce the possibility of introducing false positive detections into our optimization framework. In spite of the fair reliability of the introduced initialization mechanism, the probable improper instantiations will be analyzed in the final 'Consolidation' stage.

The compound confidence metric introduced as a complementary criterion to the NMS process for initialization, is based on 'Detection Frequency' and 'Depth-Height Map' scores. It is intuitively known that more overlapping detections in the neighborhood of a detected target, implies a higher probability of a true positive in that region. Hence the number of suppressed overlapping detections in NMS process is considered as 'Detection Frequency' and utilized as a confidence measure for the detection. The second criterion for the confidence score is based on 'Depth-Height Map', which implies the probability of a correct detection according to the relevancy of detection size and position in the frame.

We model the 'Depth-Height Map' by a first order polynomial, which describes the geometry of the scene, *i.e.,* the relationship between the average height of the pedestrians with their foot location in the frame (h, y_f). An algebraic least square method is utilized to fit the first order model to the data. The model can be estimated based on sample ground truth-ed frames. Otherwise a weighted least square estimation may be utilized to recursively calculate and update the model on the fly, given the detection results. As-

sume a Depth-Height model $\{(p_1, p_0)|\ h = p_1 \cdot y_f + p_0\}$ is estimated for the scenario, which represents the scene geometry. Then a confidence score is defined for every detection, as below:

$$C = \exp\left(k_a \cdot \left(\frac{h_p - h}{\min(h_p, h)}\right)^2\right)\ ,\ k_a = \ln(\alpha), \tag{4.12}$$

where h is the detected height of the target and h_p is the expected height based on the foot location ($h_p = p_1 \cdot y_f + p_0$). α determines the confidence value at half or double size of expected Height. We have set $\alpha = 0.1$ in our system, which results in $C = 0.91$ for a 20% height deviation with respect to the model and $C = 0.56$ for a 50% height deviation. The introduced 'Depth-Height Confidence' score is used for two purposes in our system:

- Defining cluster confidence used for instantiation of new target clusters,
- Defining track confidence used for track consolidation (described later),

The non-associated members, that are not assigned to any clusters, are constantly monitored during the optimization. As soon as the number of non-associated members with a minimum confidence exceeds a specific limit in 'n' consecutive frames, the NMS process is applied on non-associated members. If a minimum detection frequency and Depth-Height confidence is attained, new clusters will be instantiated during the clustering process.

Track Consolidation

Following a single round of clustering optimization throughout the whole video sequence, a good estimation of target tracks is acquired by the system. A final post-processing stage is required to distinguish reliable tracks among the whole set and remove spurious ones, while consolidating the incomplete tracks.

It is well known that the geometry of the scene is a strong notion in video surveillance applications, which can be leveraged to improve the system performance. For instance the knowledge of Entry/Exits in the scene is a useful notion to improve the reliability and efficiency of the system. More specifically when a pedestrian walks out a gateway or the vision field, the system

should logically stop searching for it, given the gateways are known. These are the type of information that humans use for their daily visual inspections. Hence equipping the system with such information is potentially expected to improve the system reliability. We believe taking advantage of the critical knowledge about the scene geometry is an important step towards obtaining reliable video surveillance systems.

In this work, we have used the information of Entry/Exit areas to define the completeness or integrity of the tracks. The intuition behind this definition is that, given an infinite video stream, a track can only start and finish in a gateway. This apparent criterion is utilized in our system to distinguish the incomplete tracks. These tracks are the main candidates for post-processing in the 'Consolidation' stage. The scene Entry/Exit areas can be the frame boundaries or the building doors, gateways, *etc*., which are defined by a set of bounding boxes (*c.f*. figure 4.4). Since the focus of this thesis is on video surveillance with static cameras, Entry/Exit areas are manually defined as one of the camera calibration steps during the system installation. Automatic detection of Entry/Exit areas based on the flow of the crowd is a possible extension to this work. However we don't consider this manual setting as a limitation in the proposed system, since it is a one-time simple adjustment. Furthermore for the PTZ cameras, the Entry/Exit areas can be updated with the camera movements, based on the known camera parameters and motion.

By performing the integrity test, incomplete tracks whose head or tail are not consistent with the scene gateways are recognized. Some probable situations in which incomplete tracks may emerge in the system, are:

- occlusions by static scene occluders or other moving targets may cause missed detections or confusion over an interval,
- when two closely moving targets split at some point within the scenario, and the system instantiates a new track for the newly emerged non-associated members,

The incomplete tracks should be resolved, mainly through joining to other extracted tracks, if their confidence scores (defined later) are above a threshold. This is due to the fact that all the reliable (high confidence) detections in the system have been already associated with clusters/tracks during the

Figure 4.4: Manually configured Entry/Exit data for the scenario (PETS2009, S2L1, V01)

data association stage.

We introduce two types of confidence scores for the tracks in our system, namely a 'Total Confidence Score' and a 'Distinct Confidence Score', which are utilized in consolidation process. 'Cluster Confidence' scores of a target along its whole track, is the common basis for defining both track confidence measures. The 'Cluster Confidence' is characterized by the average 'Depth-Height Confidence' of the cluster members in a frame. Assume $\widehat{C_j}(i)$ is the Depth-Height confidence of cluster center related to the target j in frame i. Then the 'Total Confidence Score' of the target track T_j is defined as:

$$S_{Tot}(T_j) = \sum_{i=H}^{T} \widehat{C_j}(i), \tag{4.13}$$

where 'H' & 'T" stand for 'Head' & 'Tail' of the track T_j in the video sequence.

On the other hand, the 'Distinct Confidence Score' of a target track T_j characterizes the independent confidence score of the track, *i.e.,* the amount of non-overlapped clusters that support the track. In other words the more a target is overlapped with other targets throughout the track length, the lower will be its 'Distinct Confidence Score'. In the light of above the 'Distinct

Confidence' is defined as:

$$S_{Dis}(T_j) = \sum_{i=H}^{T} \widehat{C_j}(i) \cdot \left(1 - \max_{k(\neq j)} O\left(T_j(i), T_k(i)\right)\right), \tag{4.14}$$

where $O\left(T_j(i), T_k(i)\right)$ represents the overlap of the target j with any other target $k(k \neq j)$ in frame i. Overlap of two targets (T_j, T_k) in a frame is defined as:

$$O(T_j, T_k) = \frac{\mathbf{area}\left[intersect(T_j, T_k)\right]}{\mathbf{area}[\min\left(T_j, T_k\right)]}, \tag{4.15}$$

In Equation (4.14) we have negligently assumed that $\max_{k(\neq j)} O\left(T_j(i), T_k(i)\right)$ represents the true overlap of the target j with all other targets in the frame. This is predominantly a true assumption in sparse crowds, where the target is not usually overlapped by more than one other target.

Duplicate tracks that are majorly overlapped throughout their whole length, demonstrate a very low 'Distinct Confidence Score'. Such spurious tracks might be generated in the clustering optimization stage, as well as the consolidation stage (discussed below). Hence the counterfeit tracks are recognized based on the proposed metric and removed from the system.

Our proposed consolidation framework, is a light-weight rule-based system which works based on the introduced metrics. Consider that the confidence scores deliberately contain a notion of tracks' lengths, as they are not normalized by the length of tracks. The tracks are resolved one at a time, in descending order of the 'Total confidence score' (S_{Tot}). Furthermore at each round of consolidation, the tracks with a low 'Distinct Confidence Score' ($S_{Dis} < \varphi$) are killed, where the threshold φ is manually set in the system. In the first step the integrity of each track is checked and the inconsistent borders (tracks' incomplete 'Head' and 'Tail') are determined. Then the incomplete boundaries of the tracks are completed by other tracks or through extension to the closest gateways in the scene. Such extensions can only occur over a few frames, due to the possibility of missed detections for the partially visible targets at gateways. A high-level flowchart of the track consolidation process is presented in figure 4.5.

The sequential rule-based consolidation process for incomplete tracks is accomplished in descending order of their total score S_{Tot}, as described be-

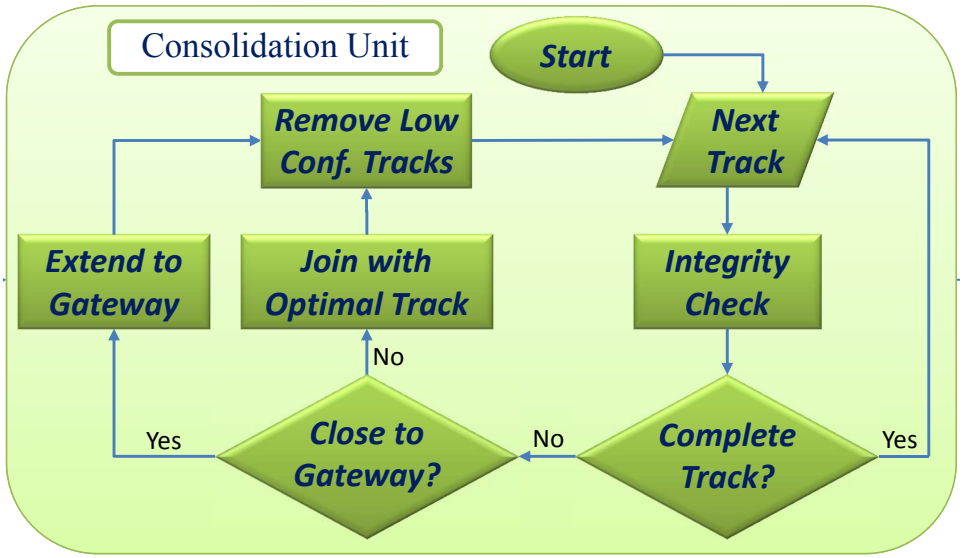


Figure 4.5: High-level Flowchart of the track consolidation process

low. We note that the set of incomplete tracks may change during the process due to the 'Split' and 'Join' operations.

- The distance of the track's incomplete boundary to the closest gateway, is estimated based on the target velocity at the boundary frame.
- Given the boundary is close enough to a gateway, the track will be smoothly extended to the specified gateway.
- If not extended, the track is completed by joining to other proper tracks through replacing or copying the track piece (described below).
- The distinct score S_{Dis} of the remaining tracks are recalculated and low confidence tracks are removed from the system.

We define a 'Correspondence' loss function for joining two tracks, to assure a smooth transfer among them. Hence the most homogeneous track with the highest similarity in position, motion and size of the target is identified for completing the specified boundary. The intuition behind joining tracks is the fact that moving targets in the scene may join or split among the video. This may result in estimation of incomplete tracks during the clustering optimization. The proposed 'Correspondence' loss function is similar in

spirit to the temporal terms of the clustering loss function (*c.f*. Equation (4.6) in section 4.3.1). Suppose $Y_k^{t_0}$ is the incomplete boundary of track k at time t_0 and $Y_i^{t_i}$ is a joining point on track i at time instant t_i. Then the energy loss for connecting $Y_k^{t_0}$ to $Y_i^{t_i}$ is defined as:

$$L_{Cor}(k,i) = \lambda_1 \left\{ \| f_S(Y_k^{t_0}) - f_S(Y_i^{t_i}) \|_2^2 \right\} + \cdots \\ \frac{\lambda_2}{3} \left\{ \| f_P(V_k^{t_0} - V_{tr}) \|_2^2 + \| f_P(V_{tr} - V_i^{t_i}) \|_2^2 + \| f_P(V_k^{t_0} - V_i^{t_i}) \|_2^2 \right\}, \quad (4.16)$$

where $\left(f_P(V_k^{t_0}), f_P(V_i^{t_i})\right)$ are the target speeds on tracks (k,i) at time instants (t_0, t_i) and $f_P(V_{tr})$ defines the transition speed between the joining points of the two tracks at (t_0, t_i):

$$V_{k(/i)}^{t_0(/t_i)} = f_P(Y_{k(/i)}^{t_0(/t_i)}) - f_P(Y_{k(/i)}^{t_0(/t_i)\pm 1}),$$

$$V_{tr} = \frac{(f_P(Y_k^{t_0}) - f_P(Y_i^{t_i}))}{|t_0 - t_i|}, \quad (4.17)$$

The first term in the 'Correspondence' loss function (Equation (4.16)) addresses the scale constancy constraint, which ensures similar target sizes. The second term assures the adjacency of the connection points and smooth transfer between the two tracks. The variable V_{tr}, encodes the relative position of the joining points on the tracks in terms of a transfer velocity. The proximity of the joining points on the tracks are not directly encouraged through the loss function. The reason is that the optimal connection point on a track, is not always the closest point. This is due to the fact that confusion of tracks occurs in occlusion situation, where there is not enough clarity about the accurate target locations. Consequently it is reasonable to look for the optimal joining point on each track, which is usually before the time an occlusion starts. Hence the solution to our problem is the optimal track T_i and the optimal joining point t_i on it, which minimizes the 'Correspondence' loss function for completing the track T_k:

$$(\psi, t_\psi) = \arg\min_{i,t_i} L_{Cor}\left(T_k, T_i(t_i)\right), \quad (4.18)$$

where ψ is the index of the optimal track T_ψ and t_ψ is the best connection

point on T_ψ which results in a smooth transfer between the identified tracks. We look for the optimal joint on each track within a limited interval in proximity of the incomplete boundary of track k.

The proposed energy minimization and motion terms of the correspondence energy loss in Equation (4.16), has some common ground to the energy minimization framework of [MRS14]. However, [MRS14] applies the motion model within a non-convex minimization problem with a high computational cost, for data association and extraction of target tracks. While at this stage of our method, the high confidence detections are already associated to tracks. Hence, the proposed correspondence energy minimization deals with the few incomplete boundaries of the established tracks rather than the numerous detection results. The minimization process explores best combination of the established tracks with negligible cost, in order to form complete tracks.

Following the recognition of the optimal track for completing the boundary of track k, we propose two strategies for joining two tracks, namely 'Join by Copy' or 'Join by Replace'. We use the 'Distinct Confidence Score' (*c.f.* Equation (4.14)) of the remaining part of track T_ψ to decide which joining strategy is suitable. More specifically, if $S_{Dis}(T_\psi) < \varphi$, then we apply a 'Join by Replace' strategy, in which the track part will be removed from T_ψ and attached to T_k (The remaining T_ψ will be removed later due to the low distinct score, S_{Dis}). Otherwise the mentioned track piece will be copied to T_k and form a duplicate track in the copied section of the track. The intuition behind this proposal is based on the fact that overlapping targets that are moving together within the scenario, may split at some point. The constant threshold φ is empirically set to $(\varphi = 20)$ in our system. However, our experiments show that any value in the range $\big(\varphi \in [15\ 50]\big)$ performs well and provides a similar performance.

4.4 Experiments

Evaluation of the system has been performed in terms of detection accuracy on one of the the most challenging datasets publicly available (sparse scenarios from PETS-2009-S2L1). The improvement over a state-of-the-art

pedestrian detector (ACF detector [DABP14]) has been presented. The tested scenario includes about 15 low resolution pedestrians, entering and exiting the scene at various time instants. Several occlusion events (more than 40 cases, some of which last for up to 15 frames) occur during the video sequence. Occlusions are due to the scene occluders or the inter-person occlusions, sometimes among pedestrians with similar appearances. The proposed framework demonstrates a robust behavior to the occlusion situations, while it is able to preserve the target identities in most of the the cases among the scenario. The overall system precision measured over the entire video sequence ($^{tp}/_{(tp+fp)}$) is increased by about 15% with respect to the state-of-the-art. This improvement is due to the elevated system stability in occlusions and the large reduction in false positive detections. Further quantitative results are provided in this section to substantiate the robustness of the proposed method through the numerous occlusion events of the tested scenario. A video result of the proposed system is also available on 'youtube', for visual demonstration [1]. The system is adapted to various number of targets along the sequence, through the online mechanisms for updating the active clusters per frame. As shown in Figure 4.2 and discussed at the end of section 4.3.1, entering and exiting targets are detected through the process and their relevant clusters are updated. Hence the system is aware of the targets number in the scene, through updating the active clusters per frame. The evaluation results on the system counting accuracy is presented later on.

A MATLAB implementation of the proposed framework on a '2.4GHz, Intel Core-i7', performs at a rate higher than 125 frames per second. This rate is calculated based on the required time for clustering optimization and track consolidation, regardless of the detection time. Hence according to the efficiency of the ACF detector, the whole system presents a real-time performance.

Figures 4.6 and 4.7 demonstrate sample outputs of the proposed framework in various major occlusions, for visual purposes. The numbers above the bounding boxes demonstrate the number of cluster members for each target. This is equivalent to the detection frequency discussed in section 4.3.2 and represents the number of the pre-NMS ACF detections which are asso-

[1] Occlusion Handling in Multiple Target Detection and Tracking

ciated to the cluster. We notice that in some frames the system represents a bounding box with '0' members. Such situations arise when the ACF detector does not report any relevant detection in that area, while our system is still able to localize the target, thanks to the temporal terms of the clustering loss function.

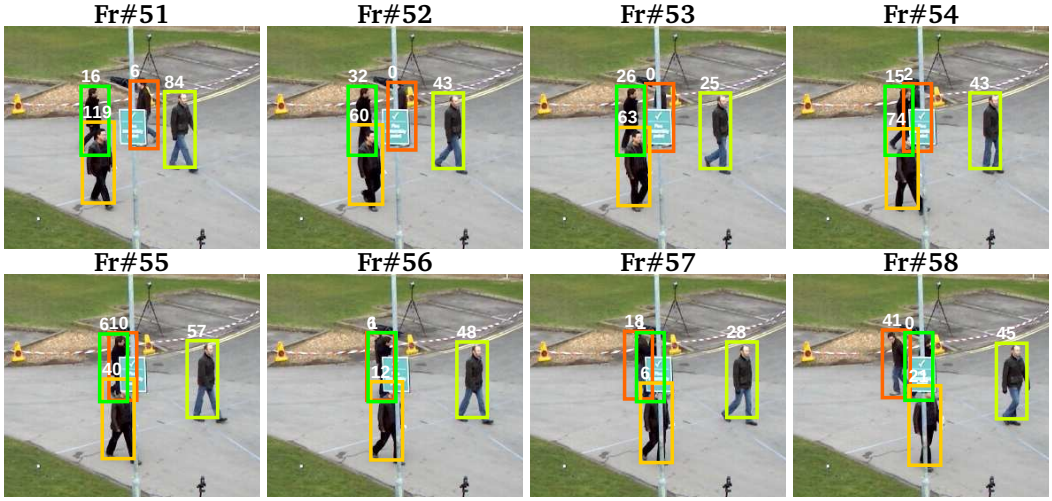


Figure 4.6: Sample results on (PETS2009, S2L1, V01), The number of cluster members are shown above the bounding boxes

The detection performance is evaluated based on PASCAL criterion [PBE+06] which considers a detection to be correct (true positive) if the overlap of the detected and ground-truth bounding box (intersection over union) is greater than a sufficient threshold (typically 0.5). We have used three different thresholds ($Thr = 0.25, 0.5, 0.75$) for comparing the performance of our system against the ACF detector. For lower thresholds, a higher performance is reported, since a smaller overlap is accepted as a correct match between the detection and ground truth.

For $Thr = 0.5$, the overall system precision $tp/(tp + fp)$ over the whole sequence), is increased from 83% in ACF detector to 96% in our proposed system, due to the large reduction of false positives. For $Thr = 0.25$, the overall precision of the ACF detector is 84.7%, while our system demonstrates 99.3% overall precision! This is due to the fact that our system keeps the track of targets among the video, although in some frames the bounding box may not be completely localized and the overlap with the ground-truth falls below 50%. Such detections are reported as an unmatched detection

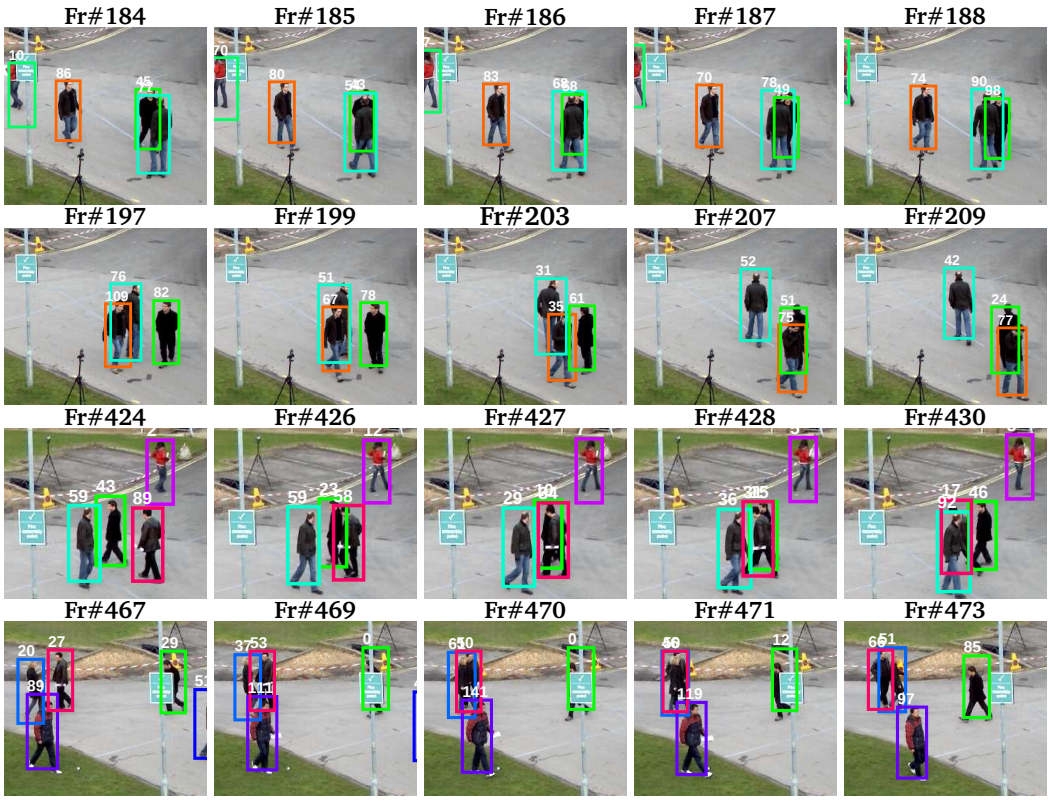


Figure 4.7: Sample results on (PETS2009, S2L1, V01), The number of cluster members are shown above the bounding boxes

with $Thr = 0.5$, while they are considered as a true match with $Thr = 0.25$.

The ROC curves of figure 4.8 for $Thr = 0.25$ and $Thr = 0.75$ clearly demonstrate the improved performance of our system as compared to the ACF detector. As shown in the curves, for $Thr = 0.25$ we get a 95% performance at $fppi = 0.04$, which is far above the ACF performance curve. However for $Thr = 0.5$ we get a 92% performance at $fppi = 0.22$, still above the ACF curve, but the trend of the ROC curve shows that we will have a much higher detection rate at higher fppi values, compared to the ACF performance which saturates at 91%.

We also evaluate a 'Counting Error' rate in the system, which is defined as the difference of the number of ground truth and detected targets per frame. The counting error was decreased from 14.5% in ACF detector to 3.8% in our system, *i.e.,* the number of targets per frame can be reported smoothly and reliably.

We notice that during the experiment one identity switch occurs between

$Thr = 0.25$

$Thr = 0.50$

$Thr = 0.75$

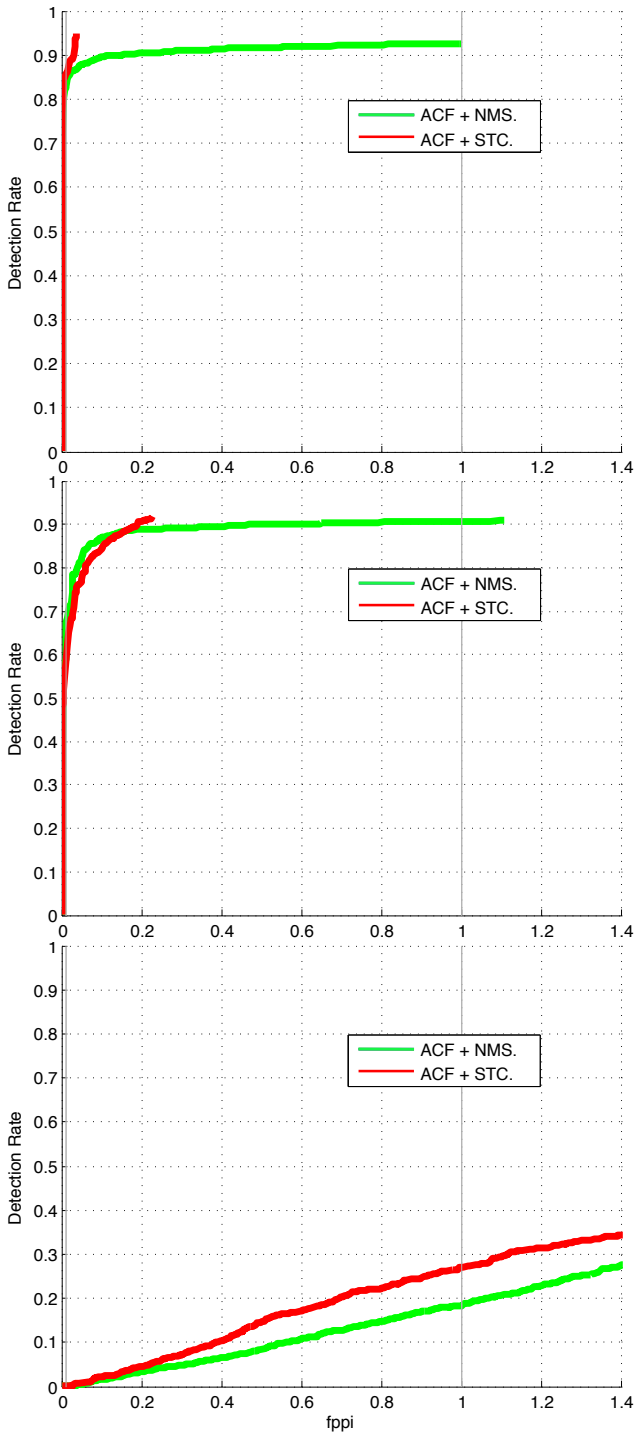


Figure 4.8: ROC curves for the two methods (ACF + NMS. *vs*. ACF + Spatiotemporal Clustering - STC.), on PETS09-S2L1-V01 scenario with three different overlap thresholds of PASCAL criterion (25%, 50%, 75%). The horizontal axis represents the False Positive Rate.

frames #425 to #430. The current proposed cost function is defined just based on the detector output features and no appearance model is incorporated in the framework yet. In the next step, we are planning to explore how to integrate an elegant appearance model in our optimization framework in order to resolve the identity switch problem, while obtaining a more accurate localization which will improve the detection rate of the system.

4.5 Summary and Discussion

We propose an efficient spatio-temporal clustering framework to improve NMS in occlusion events and simultaneously compensate for the general detection errors, namely the missed detections and false alarms. Effective occlusion handling in the proposed framework is a direct consequence of the system awareness about the number of existing clusters/targets in the frame. Spatial and temporal data association is combined in a principled framework, which entails consistency in the motion and scale of the tracked targets across frames and provides an efficient closed form solution for the problem. While the proposed cost function helps maintaining a smooth track for every target, it is also able to handle occlusion situations when the targets have different scales or motion directions. Moreover, we have taken advantage of the notions 'Depth/Height Map', the 'Scene Entry/Exit' and an 'Overlap Matrix' to post-process the results of the optimization and consolidate the estimated tracks within a light-weight rule-based system. The track consolidation module removes spurious tracks with low confidence and takes care of integrating the incomplete tracks which don't start or end at Entry/Exit areas.

The proposed system optimizes a track for each target through a clustering framework, while it realizes new pedestrians entering the scene. Hence a new cluster is defined for every emergent target within the frames, which is constantly updated along the video sequence. We suggest to use standard NMS along with a specifically designed confidence score to instantiate new clusters for new entrant targets. The above mentioned clustering cost function along with the proposed scheme for cluster instantiation, seems to be adequate enough to mitigate the need for Foreground/Background modeling towards reducing the false positives. The proposed framework significantly

improves the detection performance in sparse crowds, by removing most of the false positive detections.

The proposed framework demonstrates a high capability for occlusion handling within a low resolution context, given a proper pedestrian detector is utilized. We use an efficient state-of-the-art pedestrian detector [DABP14] with a full-body pedestrian model in the first stage of our system. Hence the whole system is computationally very efficient and provides a real-time performance on a standard CPU. As mentioned above, the proposed framework is suitable for processing both high or low resolution video sequences. Hence the system performance does not decrease very much with low quality videos, due to using a whole body model in detection. The above mentioned properties, including the system compatibility with the existing infrastructures and low resolution cameras, make it attractive for practical video surveillance.

Future works: One potential extension to the current framework is to incorporate the targets' appearance models in the clustering optimization and track consolidation. This is expected to improve the system performance further in occlusion situation and resolve the remaining confusions. More clearly when the motion and size information alone do not suffice to resolve the ambiguities and identity switch among the targets are probable, an elegant appearance model is expected to improve the system behaviour. We are curious to investigate how much this generalization can improve the detection and tracking performance in medium and dense crowds, scenarios like people walking in a mall or exiting a stadium. Moreover we are aware that further experiments and more thorough evaluation could help to improve and demonstrate the strength and stability of the system for practical video surveillance applications.

Chapter 5

Hardware Acceleration, Parallel Processing

5.1 Introduction

It is well known that human and primates outperform all the existing computer vision systems by almost any measure in various visual tasks such as recognition. Hence building artificial vision systems that emulate the capabilities of the brain cortex layer has always been an attractive idea. Biologically inspired research in computer vision has produced a number of promising models which replicate the lowest levels of the visual cortex for extracting visual features [WRC08]. In this primary level, visual cells extract local oriented features in various scales and orientations, to be used as fundamental cues for constructing a detailed representation of the vision field.

Different types of multi-scale and multi-Orientation filters have been proposed in the computer vision literature for extracting local oriented features [Kub95]. Among the existing models, 'Gabor' filters and 'Gaussian Derivatives' [SWP05, Dia06] are the most popular ones which simulate the function of simple cells in visual cortex. The proposed models are multi-channel filter banks, where all the channels are generated from one mother wavelet through dilation and rotation. Each filter channel represents a specific scale and orientation in 2D or 3D space.

In the last decades multichannel decomposition of images and videos have been applied in various computer vision applications. The Spatiotemporal

Oriented Energy features which we used in previous chapters for the purpose of space-time video analysis (*c.f*. Chapter 3), are a 3D extension of 'Gaussian Derivative' functions in space-time (x, y, t). Increasing application of biologically inspired features in computer vision along with their relatively high computational cost, motivates using parallel computing technologies such as GPUs. The importance of hardware acceleration in computer vision is unsubtle due to the fact that a wide range of applications in video surveillance, robotics and human computer interaction require real-time processing of a streaming video. The profoundly parallel nature of biological vision systems and neural networks is also a good support for this approach.

In a pilot plan prior to our main research, we developed a GPU engine of a 2D filter bank, to evaluate the best achievable performance for extraction of multichannel features. In this chapter we present details of a Gabor GPU kernel for extraction of the 2D multichannel Orientational Gabor features in real-time. The design of the GPU kernel mimics the initial layers of visual cortex composed of 'Simple' and 'Complex' cells. 'Simple' units perform a pattern matching with the Gabor wavelets in different scales and orientations by filtering the image. The filtering operation is conducted through convolving the image with the filter masks. Then 'Complex' units aggregate the outputs of the 'Simple' units at common orientations to provide local spatial invariance.

5.2 Related Works

Multi-scale and multi-directional processing are fundamental concepts in artificial and natural vision systems for perceiving the environment. This is due to the fact that real world objects and structures appear in different orientations and scales depending on their relative position to the observer. Hence a legitimate visual perception of the environment without considering the notion of scale is impossible. In other words a reasonable solution is required to handle the multi-scale nature of real-world objects and their dynamics. In absence of any prior information about the relevant image scales, a common approach is to build an image pyramid representing the visual information at multiple scales. In such a multi-scale representation, the coarse pyramid levels are simplified versions of the corresponding structures at fine scales

with less details. The most well known pyramids which are widely applied in various applications, are Gaussian pyramids [Bur81] and Laplacian pyramids [BA83].

Inspired by the function of the visual cortex cells, more principled approaches have been proposed to handle the multi-scale and multi-orientation analysis in a unified framework. Gabor filters and 'Gaussian Derivatives' are the most popular multi-scale orientational filters in the field of computer vision. The Gabor model for the response of simple cortical cells was first proposed by Marcelja in 1980 [Mar80]. However the mathematical functions are often credited to Dennis Gabor in 1946, who supported their application in communication systems [Gab46]. Gabor filters have been widely applied in various computer vision applications, such as facial recognition [LW03], iris recognition [MWT02], fingerprint recognition [AJP00], texture segmentation [JF91], *etc*. The second model of cortical receptive fields, known as Gaussian Derivatives, was first proposed by Young in 1986 [You86]. Due to their natural bond with Gaussian pyramids, Gaussian Derivatives have been widely applied for multi-scale processing of images and videos [Lin94, Bur81, BA83].

Gabor model is a set of complex valued functions, characterized as the product of a Gaussian kernel and sinusoidal plane waves expressed by complex exponential. However Gaussian Derivatives are real valued functions represented as product of a Gaussian kernel and the generalized Hermite polynomial. Hermit polynomials are generated through partial derivations of the Gaussian kernel [YL01]. Gaussian Derivatives are also known as the n^{th} Derivative of Gaussian (DoG).

Gaussian derivatives and Gabor functions have adjustable spectral bandwidths and can be rotated to construct a filter bank. Orientation in Gabor filters is an explicit parameter which is directly tunable and makes the filter intrinsically steerable [WC05]. While the DoG filters are synthesized from linear combinations of basis functions [FA91]. For visual demonstration, the basis functions of 2^{nd} Derivative of Gaussian filters are shown in Figure 5.1. The steerable filters used in previous chapters are complex quadrature pair filters constructed based on DoG functions and their Hilbert transform, which were first introduced by Freeman [FA91].

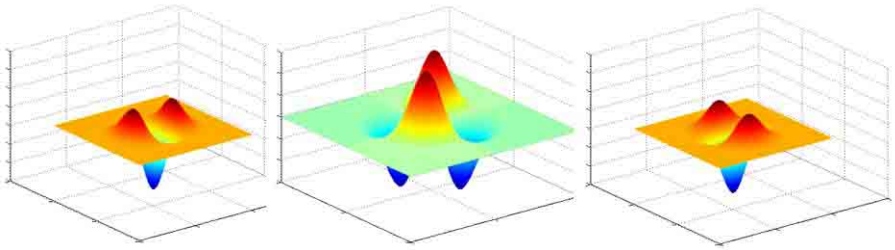


Figure 5.1: Basis functions of 2^{nd} Gaussian Derivatives

Many variants of multi-scale and multi-orientation transforms, have been developed in the last two decades. To name a few, 'Dual-Tree Complex Wavelet Transform' [Kin99, SBK05], 'Pyramidal Dual-Tree Directional Filter Bank' [NO08, NO06], 'Contourlet' Transform [DV05] and 'Curvelet' Transform' [EJCY06, NC10]. The scaling factor and bandwidths of the filters in all of these multi-channel filter banks, are designed to efficiently cover the whole spacial frequency domain (*c.f*. Figure 5.2). Hence all the filter banks provide a rich and efficient description of the visual domain. However only some of them provide a perfect reconstruction scheme for reverse transform. For instance the Gabor wavelets are not orthogonal and do not produce a perfect reconstruction scheme. They are highly redundant due to unlimited possible number of orientations and produce an overcomplete representation of image. We refer interested readers to [VO08], for further comparative details about the above-mentioned multi-channel transforms.

Due to the wide application of Gabor filters in computer vision and their intensive computations, we conducted our pilot plan for parallel processing on Gabor filters. To obtain a Gabor filter bank with certain number of scales and orientations, all the filters are generated from one mother wavelet by dilation and rotation. The two-dimensional Gabor filter in spatial domain, is a Gaussian kernel modulated by complex sinusoidal plane waves:

$$g(x,y) = \frac{e^{j(2\pi x'/\lambda+\psi)}}{2\pi\sigma_x \cdot \sigma_y} . \exp\left(-\frac{1}{2}(\frac{x'^2}{\sigma_x^2} + \frac{y'^2}{\sigma_y^2})\right), \tag{5.1}$$

Where λ is the wavelength of the sinusoidal factor and ψ is the phase offset. Moreover σ_x and σ_y represent the standard deviation of the Gaussian envelope along the x and y directions. Hence $\left(\sigma_x\,,\,\sigma_y\right) = (\sigma\,,\,\sigma/\gamma)$, while σ is

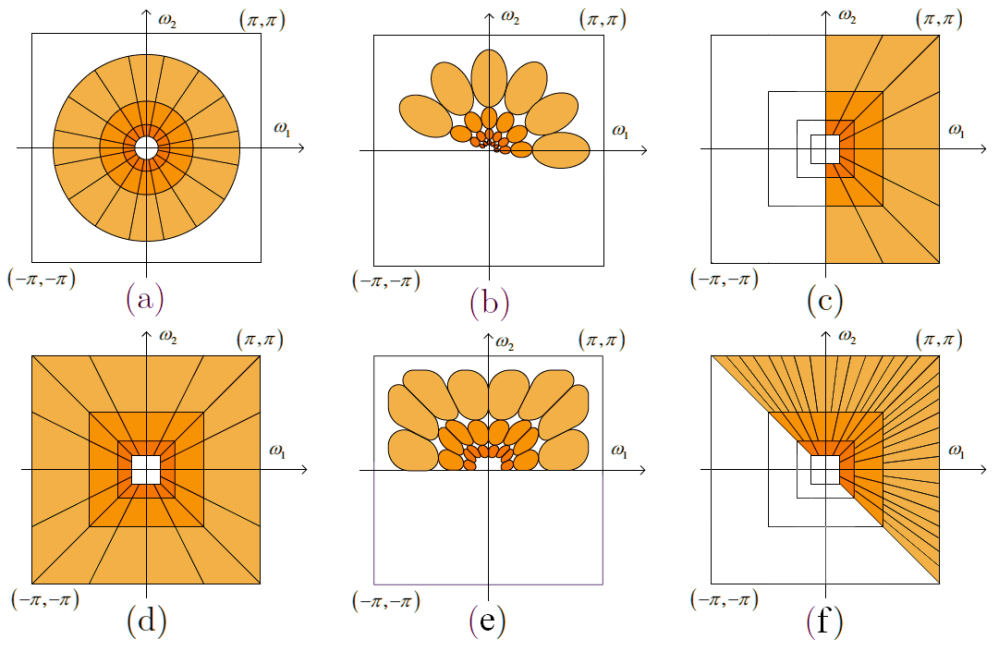


Figure 5.2: 2D Frequency response of Multi-Channel Wavelet Transforms: (a) Steerable pyramid (2^{nd} DoG): $(N_S, N_\theta) = (3, 8)$, (b) Gabor Wavelets: $(N_S, N_\theta) = (4, 6)$, (c) Dual-Tree Complex Wavelet Transform: $(N_S, N_\theta) = (3, 8)$, (d) Contourlet transform: $(N_S, N_\theta) = (3, 8)$, (e) Pyramidal Dual-Tree Directional Filter Bank: $(N_S, N_\theta) = (3, 8)$, (f) Uniform Curvelet Transform: $(N_S = 3, N_\theta = 4, 8, 16)$.

the total standard deviation along the filter orientation and γ indicates the spatial aspect ratio relevant to the ellipticity of the Gabor support. x' and y' are functions of the filter orientation θ, as defined below:

$$\begin{pmatrix} x' & y' \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix} \cdot \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{5.2}$$

Filtering is carried out across multiple channels at different frequencies ($f = 1/\lambda$) and orientations (θ), through spatial domain convolution [HX05, LW02]. Gabor filters are non-separable filters by nature, thus requiring a 2D convolution for filtering operation. The order of complexity in spatial convolution is $O(M^2.N^2)$, where (M, N) are the sizes of image and filter mask. Many research efforts have been made to improve the computational complexity of Gabor filtering [NNPT98, AWT04, AWST05a]. Due to separability of the Gabor functions in horizontal and vertical directions, some studies propose to reconstruct the image along the filter orientation [AWST05b]. Hence

following the reconstruction process, the separable functions may be applied rather than the 2D Gabor filters. However such methods require a resampling process on the image for every single orientation. This implies a costly 2D interpolation on the whole image per filter channel, which increase the computational complexity of the system. Some other studies suggest to perform the convolution operation in frequency domain [ATB09]. In frequency domain convolution, Fast Fourier Transform (FFT) of the image is multiplied by the FFT transformed Gabor filter. Then the multiplication result is converted back to spatial domain using the inverse FFT. Performing filtering process in the frequency domain reduces the computational complexity to an order of $O(M^2.\log N)$. One issue in this approach is that the generic FFT formulation is limited to signals with an even length ($2n$). Furthermore the FFT based methods require large memories for keeping the intermediate results among the process [ATB09]. As described in the next section, optimizing memory transactions is a critical designing issue in GPUs, which can affect the total performance. Hence in this work we use a spatial domain convolution for designing our GPU engine.

5.3 Background on GPUs

GPUs are inexpensive parallel processors that have been widely employed in many applications as powerful coprocessors. The power of GPUs as hardware accelerators stem from their high memory bandwidth and large number of programmable cores. Thousands of hardware thread contexts execute the programs within a SPMD model(Single Program, Multiple Data). GPUs are flexible and easy to program, thanks to the high level languages and available APIs which encapsulates the hardware details. Compared to FPGA design process which requires frequent hardware corrections, modifying a GPU function is straightforward through recompiling a modified code.

In this work, we use OpenCL API within the CUDA architecture for developing a GPU kernel of Gabor filter bank. OpenCL which stands for 'Open Computing Language', is a framework for writing programs that execute across heterogeneous platforms consisting of CPUs, GPUs, and other processors. OpenCL provides access to the Graphical Processing Unit (GPU) for

non-graphical applications through the existing APIs. CUDA represents the nVIDIA "'Unified Device Architecture"'. In CUDA, the GPU is treated as a coprocessor that executes data-parallel kernels with thousands of threads. Threads are grouped into thread blocks, which share data using fast shared-memories. There is no communication among thread blocks or working groups, *i.e.,* their execution is completely independent. The coordination of the thread blocks is conducted through a large global memory, which is much slower compared to the dedicated shared memories.

Execution Model - CUDA Architecture

The CUDA architecture corresponds to the OpenCL architecture. CUDA device is built based on a scalable array of multithreaded Streaming Multiprocessors (SMs). A multiprocessor executes a thread block for each OpenCL work-group. A kernel is executed over an OpenCL 'N' dimensional grid of thread blocks (NDRange). As illustrated in Figure 5.3, each of the thread blocks is uniquely identified by a work-group ID. Moreover each single thread is identified by a unique global ID or through a combination of a local ID and the work-group ID [Ope09].

Every work-group in the grid has access to a limited high-speed shared memory. Typically, each work-group invokes hundreds of threads. Generally every 32 scalar threads are combined in a warp which is defined as a working group in the SIMD (Single Instruction, Multiple Data) model. A common strategy for GPU programming is to split the problem into blocks with similar function, operating on different parts of the information (SIMD). Hence the kernel blocks execute in parallel on distinct data.

Memory Hierarchy

Memory bandwidth is a critical issue in GPUs. The memory resources should be dedicated wisely, in order to optimize memory transactions. Otherwise data transmission might be much more time consuming than the actual process, due to massive parallel computing resources and their data transactions in the GPU. When programming on GPU, the programmer has direct access to any part of the memory hierarchy. In fact it is encouraged to design applica-

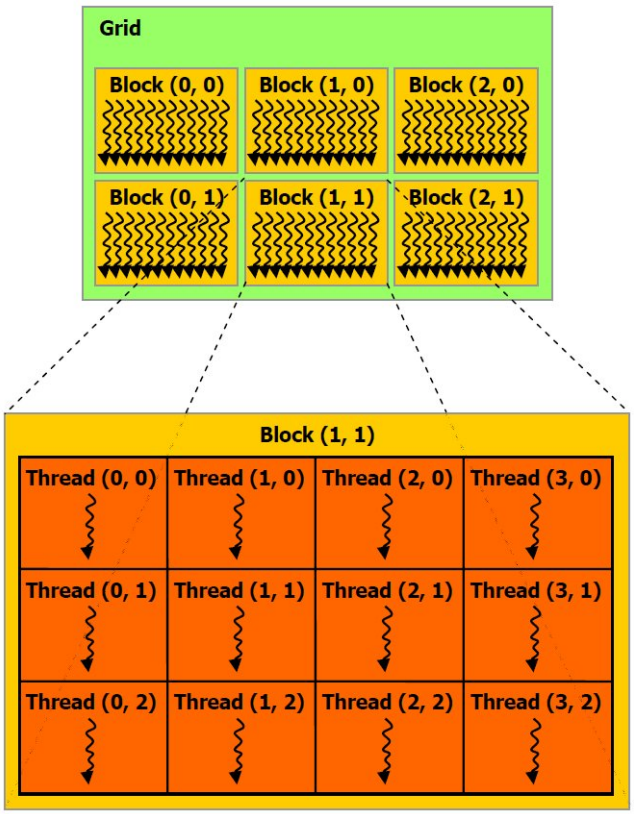


Figure 5.3: CUDA Execution Model

tion specific memory hierarchy to optimize the transactions and therefore the execution time. Three types of memory is provided in the GPU architecture:

- **Local registers**: Each GPU core has some local registers which serve as the fastest read-write locations available to the threads.
- **Shared memory**: CUDA exposes a fast shared memory region amongst all the threads in a block. Multiple blocks running on the same Streaming Multiprocessor (SM), use separate parts of a shared memory. This high-speed memory is used as an interface cache memory between the global memory and active threads, due to its higher speed and bandwidth compared to the global memory.
- **Global memory**: The GPU's global memory may be accessed by all threads and blocks. However it is several times slower than the shared memory, as it requires several hundred cycles for each data transfer. This large latency may be hidden, if the communication cost is much

lower than the computational cost. Furthermore the average transaction latency may be decreased, given the kernels are designed for optimum coalesced access to the global memory.

5.4 Technical Approach

The biological studies of the simple cells in visual cortex, indicate that the typical aspect ratio (γ) in the Gabor model is in the range of 0.2 and 1 ($0.23 < \gamma < 0.92$) [PK97]. An aspect ratio less than unit, implies an elliptical rather than a circular support for the Gabor function. In this case, a rectangular filter mask might be more suitable for representing the elliptical envelope. General trend in the literature is to assume Gabor filters as square masks, to simplify the system design specially in FPGA based processors. A direct consequence of this simplification is that a lot of practically zero coefficients in the mask, lead to extra computational cost for the system. On the other hand some non-zero coefficients may be removed from the mask which causes computational errors or inaccuracies in the estimated features (*c.f*. Figure 3.7).

To improve on these issues, we designed an adaptive Gabor engine for convolution operation in spatial domain. The proposed engine supports non-square filter masks with a variable size adaptable to different channels. In other words, the filter mask size in various orientations is adjustable to the shape of Gabor support. Hence the computational cost is minimized without sacrificing the accuracy. Optimum size of the filter mask depends on orientation of the filter, as well as the standard deviation of the Gaussian envelope. We propose to calculate the width and height of the Gabor masks (F_W, F_H) through the following equation:

$$F_W = 1 + 2\max(\,|n_s\sigma_x\cos(\theta)|\,,\,|n_s\sigma_y\sin(\theta)|\,), \tag{5.3}$$

$$F_H = 1 + 2\max(\,|n_s\sigma_x\sin(\theta)|\,,\,|n_s\sigma_y\cos(\theta)|\,),$$

$$\sigma = a\cdot\lambda \;\;,\;\; (\sigma_x\,,\,\sigma_y) = (\sigma\,,\,\frac{\sigma}{\gamma})\,,$$

where λ is the wavelength of Gabor oscillation in pixels and θ is the filter ori-

entation. (σ_x, σ_y) represent the standard deviation of the Gaussian envelope along x and y axis and γ is its spatial aspect ratio, specifying the ellipticity of the envelope. The constant integer number n_s is manually set to define the half-length of a base symmetrical mask. Furthermore the constant 'a' is a function of the filter bandwidth (BW), as defined in Equation 5.4 [WRC08], with the bandwidth specified in octaves.

$$a = \frac{k}{\pi} \cdot \frac{2^{BW}+1}{2^{BW}-1}\ ,\ k = \sqrt{\frac{\ln 2}{2}}. \tag{5.4}$$

Definition of the filter sizes in Equation 5.3 assures a minimum accuracy in feature extraction, since the filter coefficients above a certain threshold will not be discarded. The kernel is designed to calculate real and imaginary parts of the complex Gabor transform separately. Hence two separate matrices are calculated for every channel which represent the energy and phase features of the image in that specific channel. "'Feature Pooling"' with Max and Histogram methods is the last layer in the proposed Gabor engine, with a negligible additional load on the system. This layer emulates the function of "'Complex cells"' in visual cortex, by spacial aggregation of the calculated features in similar channels.

The nVIDIA CUDA framework enables us to use the Graphics Processor Unit as a general purpose computing device. We implemented our Gabor engine on nVIDIA GTX285 device (*c.f*. Figure 5.4) by using OpenCL. The nVIDIA GeForce GTX285 is comprised of 30 Streaming Multiprocessors (SM). Each SM has 8 Streaming Processors (SPs), which implies a total 240 SP cores. Furthermore each of the SP cores are deeply multithreaded. The maximum number of active warps per multiprocessor (SM) is 32, which indicates 4 warps or 128 threads per SP. As a result the number of active threads per SM is limited to a maximum of 1024. Hence GTX285 is capable of executing a maximum number of 30720 (30×1024) threads in parallel. However this processing capability is dependent on many conditions including the limitations of shared memory and local registers.

GTX285 provides 16KB shared memory per multiprocessor, organized into 16 memory banks. Furthermore 2048 local registers are provided per SP, denoting a total number of 16384 registers per SM. This implies that for acti-

Figure 5.4: nVIDIA GeForce GTX285 device

vating the maximum number of threads on a multiprocessor (1024 threads), only 16 registers can be dedicated to each thread [CUD09]. The architecture of the designed kernel along with the implementation details are discussed in the following section.

5.4.1 GPU Kernel Structure

The proposed adaptive Gabor kernel, divides the input image into blocks of 16×16 pixels for processing. Hence a preprocessing step on the input image is required to pad the width and height of the image to multiples of 16 pixels. Similarly, dimensions of the filter masks are considered as multiples of 16. This doesn't impose any limitations on the filter size, as the zero padding may be applied to any filter size. Given a non-unit aspect ratios ($\gamma < 1$), various channels of the Gabor filters have different sizes corresponding to their orientation (θ). Hence the GPU thread blocks concurrently process the same image with filters of variable sizes.

When the convolution process is finalized in all of the thread in a "'Working Group"', an efficient "'Spatial Feature Pooling"' is conducted through "'Parallel Reduction"' technique. For every block of 16×16 pixels, two types of aggregated features are provided per channel:

- Maximum of the Gabor features in the block,
- Summation of Gabor features in the block, to be used as one of the Histogram bins.

The additional computation cost for pooling is negligible compared to the convolution operation. However the abstract results provide meaningful information about the image structure at common orientations, with some level of spatial invariance. The produced integral Gabor features have been widely applied in many studies.

Processing Resources

Our kernel is executed over an OpenCL 3DRange, within a 3 dimensional grid of thread blocks. Every thread block consists of $16 \times 16 \times 1 = 256$ threads. Each thread is responsible for calculating the filtered value of a single pixel, by convolving the filter mask on a sub-image block around the pixel. The Multiplication-Addition operations of the two convolutions for real and imaginary parts of the features, are sequentially conducted in a thread. In this structure every thread block is responsible for processing 256 pixels in an image block to produce one channel of the Gabor features for all pixels. Based on the assumptions, the image is divided into a grid of blocks. Suppose there are N_x block columns in the image width and N_y block rows along the height. Also assume the filter bank has N_S scales and N_θ orientations. Then a 3DRange grid of processing blocks is organized in the GPU for parallel processing and extracting the Gabor features. Hence, the 3DRange grid dimensions S_{Grid} are defined as:

$$S_{Grid} = \{N_x, N_y, N_{Ch}\} \;\; \& \;\; N_{Ch} = N_S \cdot N_\theta \;, \tag{5.5}$$

where N_{Ch} represents the number of filter channels. For the purpose of demonstration, Figure 5.5 shows the real and imaginary parts of Gabor channels in a typical Gabor bank with 3 scales ($N_S = 3$) and 6 orientations ($N_\theta = 6$). As discussed beforehand, the size of filter masks used in our engine are variable in different channels, depending on their scale and orientation.

The proposed architecture may be described as N_{Ch} processing layers. Each hypothetical layer, consists of $N_x \times N_y$ thread blocks which apply one Gabor channel to the whole input image in parallel. Every layer use a distinct filter mask corresponding to the relevant scale and orientation of the Gabor channel. Hence all the $N_S \times N_\theta$ Gabor channels concurrently apply on a sin-

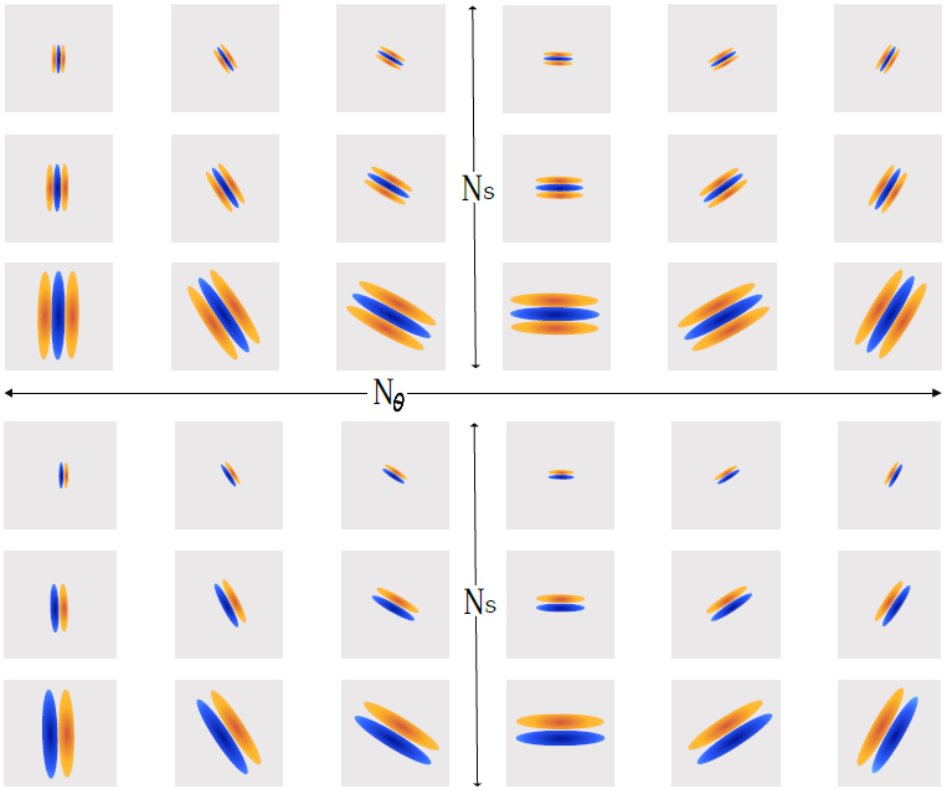


Figure 5.5: A Gabor bank with 18 channels ($N_S = 3$ and $N_\theta = 6$)

gle image within the GPU processing grid. Input frames are sequentially fed into the engine. To have a better understanding of the whole computational space inside the GPU, let's assume the physical processing resources are unlimited. Given the image width I_W and image height I_H, the total number of simultaneously active threads is:

$$S_{Global} = I_W \times I_H \times N_{Ch}. \tag{5.6}$$

However S_{Global} is much larger than the available processing cores. Hence the GPU controller manages an optimal parallel-serial scenario to execute all the thread blocks efficiently, until the whole process is accomplished. Given the local memories and internal registers are efficiently utilized, one Streaming Multiprocessor (SM) in the GTX285 is capable of simultaneously running 1024 threads or 4 image blocks of 16×16 pixels. Hence all the 30 SMs together, can potentially process 120 blocks in parallel. This is equivalent to a total $120 \times 256 = 30720$ concurrent convolution operations.

Memory Resources

One of the the most critical issues in GPU is the memory bandwidth and data transactions. Huge amount of information transfer among the processing units and GPU memories might turn to a system bottleneck. For instance, one 2D convolution operation with a typical Gabor mask of size 16×16, requires 2×256 data fetches for accessing the image and filter information in the memory. Apparently performing 30720 concurrent convolutions can put a considerable strain on the GPU memory bandwidth. Hence optimal strategies for data fetching and memory management are required to achieve good performance.

All the threads in a thread block work with the same filter matrix, for filtering various pixels of an image block. Thus fetching a filter matrix is performed through a 'Broadcast' to all of the threads in a block. Due to low speed of the global memory, the filter matrix is cached prior to broadcasting. Size of the Gabor filter masks are fairly large and the filter elements are float numbers, taking four bytes per element. Consequently, due to the limited space of the local memory, catching the whole filter mask may lead to a memory overload. To address all these technical issues, the system is designed to buffer 16 elements of the mask per transaction. Hence a coalesced access to global memory is performed to minimize the transaction delay. For proper management of the filter information in the thread blocks, the width of the masks are padded to multiples of 16. In this way every row of the filter mask is fetched in one or two transactions, depending on the size of the mask. Filter masks are fetched in 16 packs and the filtering process is conducted sequentially inside a thread.

On the other hand, every thread dedicated to processing a pixel, requires an image block around the pixel for filtering. For minimum access to the low speed global memory, every 16×16 thread block caches a larger area around the image block by coalesced accessing to global memory. Then every single thread in the working group, reads the buffered image by parallel access to non conflicting local banks. We use single channel images with 8 bit depth in this work. Thus the required cache memory for storing the image block is less than the local memory limits. For example with a filter

mask of size 32×36, each 16×16 thread block caches an image area of size $(16 + 32) \cdot (16 + 36) = 2496(B)$ in local memory. Through this strategy the access to global memory for fetching the image data is optimized by coalescing and the total number of accesses is considerably reduced. In order to calculate one channel of the Gabor features for the whole image, the total number of fetching the image data from the global memory is given by:

$$\frac{(N_x \times N_y) \times (S_B + F_W) \times (S_B + F_H)}{(N_x \times S_B \times N_y \times S_B)} = (1 + \frac{F_W}{S_B}) \times (1 + \frac{F_H}{S_B}), \qquad (5.7)$$

where S_B is the block size ($S_B = 16$), F_W is the mask width and F_H is the height of the filter mask. For $(F_W, F_H) = (32, 36)$, the source image is fetched less than 10 times and for $(F_W, F_H) = (16, 16)$, the source image is only fetched 4 times.

5.5 Experiments

For evaluation of the proposed adaptive Gabor engine, we used gray-scale VGA size (640×480) images. The engine throughput is reported by frame rate (FPS) and pixel rate (MP/S), for various fixed size filter banks. Furthermore, the throughput of the adaptive Gabor engine with variable filter sizes, is evaluated and compared against the performance of the fixed size filter banks. We finally present the speedup performance of the proposed GPU engine compared to an efficient CPU implementation.

Table 5.1 presents the throughput of a 12-channel Gabor bank, comprised of 4 orientations and 3 scales, with fixed size kernel.

Table 5.1: Throughput of fixed size 12 channel Gabor engines *vs*. the kernel size

Filter Size	7×7	11×11	15×15	19×19	25×25	35×35	45×45
FPS	59	42	33	23	14	6	4
MP/S	18.3	13.1	10.2	7.3	4.4	1.9	1.4

For evaluation of the adaptive Gabor engine, the filter bandwidth is set to one octave ($BW = 1.0$) and the aspect ratio of the Gaussian envelope is fixed on $\gamma = 0.8$. The optimum mask sizes for different channels are calculated by the engine, based on Equations 5.3 and 5.4. The mask size depends on the standard deviation of the Gaussian envelope and the filter orientation. Lower

frequencies indicate larger wave lengths, which leads to a larger Gaussian envelope and filter mask. Table 5.2 demonstrates the mask sizes of 16 different channels in a Gabor bank with 4 Orientations and 4 Scales.

Table 5.2: Filter mask sizes (F_W, F_H) *vs.* filter channels (λ, θ)

θ / λ	$\lambda = 2$	$\lambda = 4$	$\lambda = 6$	$\lambda = 8$
$\theta = 0$	$(9, 11)$	$(15, 19)$	$(23, 27)$	$(29, 35)$
$\theta = \pi/4$	$(7, 7)$	$(13, 13)$	$(19, 19)$	$(25, 25)$
$\theta = \pi/2$	$(11, 9)$	$(19, 15)$	$(27, 23)$	$(35, 29)$
$\theta = 3\pi/4$	$(7, 7)$	$(13, 13)$	$(19, 19)$	$(25, 25)$

Performance of the adaptive Gabor engine is evaluated with a 12-channel Gabor Bank in order to be comparable against the results of Table 5.1. Four orientations and the three first scales in Table 5.2 ($\lambda = 2, 4, 6$) are used to build up the Gabor channels. The throughput of the adaptive engine with variable kernel, is measured as 30 VGA frames per second or equivalently 9.2 MPixels per second. This is comparable to throughput of a 15×15 fixed size kernel, as shown in Table 5.1, in spite of using mask sizes up to 23×27 in specific channels. As discussed earlier in this chapter, using fixed size filter banks may lead to truncation of useful mask coefficients on some channels. On the other hand, it may cause unnecessary computational cost for zero-multiplications on other channels.

Figure 5.6 illustrates three optional filter masks for a Gabor function at orientation $\theta = \pi/6$. The largest filter mask is a non-optimal choice due to many zero coefficients in the mask, which cause unnecessary computational cost. The smallest mask is also not a good choice, as truncating the useful coefficients may lead to inaccurate features. The middle rectangle is considered as an optimum mask in the proposed solution, which provides accurate results with minimum cost.

In the last experiment, the speedup performance of the proposed GPU engine is evaluated. To this end the adaptive GPU engine is compared against a CPU based implementation of the proposed Gabor bank. The CPU engine is designed based on the standard OpenCV library with an efficient C++ code. Evaluation of the CPU engine has been conducted on an Intel Dual Core CPU 6400 @ 2.13 GHZ. Table 5.3 demonstrates the execution time of the CPU and

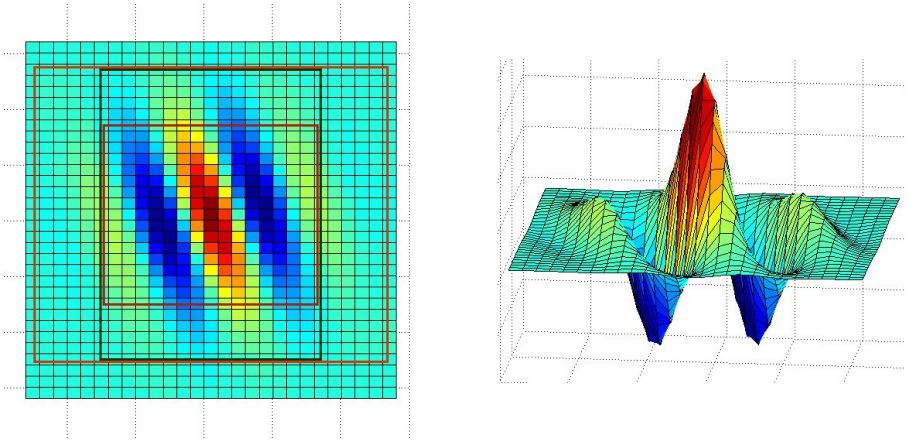


Figure 5.6: Optimum filter mask for a Gabor function at orientation $\theta = \pi/6$

GPU engines with various kernel sizes, for filtering a single VGA size frame. As demonstrated in the table, the speed up gain achieved by the GPU, grows with the size of filter masks.

Table 5.3: GPU-CPU Performance Comparison for the 12 Channel Gabor Filter Bank

Filter Size	7×7	9×9	11×11	13×13	15×15	17×17	19×19	25×25
CPU (ms)	3936	5436	7320	9564	12192	15372	18600	30780
GPU (ms)	16.7	20.0	23.3	26.6	29.8	34.5	42.0	69.6
speedup	235	271	314	359	409	445	442	442

5.6 Summary and Discussion

Increasing application of the multi-scale oriented features in computer vision and their relatively high computational complexity, encourages using parallel computing technologies. In this chapter, we presented a high throughput GPU engine for extraction of 2D multichannel Gabor features with real-time performance on VGA frame sizes. The proposed adaptive Gabor engine supports 'Non-Square' kernels, while adjusting the kernel size per channel, for efficient and precise estimation of the multi-channel features. The design of the GPU kernel follows the parallel structure of the initial visual cortex layers, composed of 'Simple' and 'Complex' cells. Simple units apply various Gabor filter channels with different scales and orientations to the input image. Complex units aggregate the extracted features by the Simple units at

common orientation and scale. Aggregation is performed through a feature pooling with 'Maximum' and 'Histogram' methods. The aggregated features provide abstract information with local spatial invariance, about the orientational structure of the input image.

This GPU accelerator was developed prior to our main research, to evaluate the speedup performance for extraction of the multichannel features. The achieved gain by the proposed engine, indicate that GPU acceleration on the multi-channel Spatiotemporal Oriented Energies (utilized in other chapters), may obtain real-time performance. Further research is required to investigate the effect of filter size variations, non-square filter masks and elliptic Gabor functions ($\gamma < 1$), in the performance of applications such as detection and recognition. Application specific metrics should be employed to evaluate the flexible kernel, towards obtaining optimal and efficient solutions.

Chapter 6

Summary and Conclusions

This dissertation contributed mainly towards occlusion handling in pedestrian detection and tracking for video surveillance applications. The proposed methods are potentially applicable to general visual tracking as well. However the established frameworks have been only evaluated on challenging video sequences of pedestrians and crowds.

Chapter 1 of the dissertation served to motivate the importance of the occlusion handling problem in video surveillance and visual tracking, introduce the research gaps addressed by the thesis and overview the thesis contributions and outline.

Chapter 2 reviewed some representative methods in visual tracking with single camera, along with their recent evolutions in the literature. Furthermore the relevant occlusion handling approaches for two typical families of tracking systems were discussed. The limitations and shortcomings of the existing approaches were elaborated to prepare a proper context for the research problems addressed by this thesis.

Chapter 3 presented an occlusion analysis framework for template tracking systems, based on Spatiotemporal Oriented Energy (SOE) features of the targets. SOE features provide a rich description about motion dynamics of the targets, which are applied for the purpose of occlusion modeling. Background preliminaries on SOE features are discussed prior to presentation of the main framework.

Chapter 4 proposed a spatio-temporal clustering framework for occlusion handling in the context of multiple tracking by detection, based on an off-line

trained full-body model of pedestrians.

Chapter 5 discussed parallel processing and hardware acceleration based on GPUs for speeding up the intensive calculations. A GPU engine for extraction of 2D multi-channel features was developed to demonstrate the speedup gain by the proposed hardware accelerator.

The major contributions of this thesis have been made in three areas: **(i)** Occlusion handling in template-based visual tracking; **(ii)** Occlusion handling in detection-based tracking; and **(iii)** GPU-based hardware acceleration. In the following sections we summarize the contributions in each area and suggest possible extensions and future research directions.

6.1 Occlusion Handling in Template-based Tracking

Chapter 3 presented an occlusion handling framework based on motion dynamics of the targets. Motion dynamics are described through the multi-channel Spatio-temporal Oriented Energy (SOE) features. The system provides bases for protecting the target model against corruption and improving the tracking performance in occlusion situation. The proposed method has been evaluated on a single target tracker, for the proof of concept. However the proposed method is not limited to single target tracking and may be applied to any template based tracking system, in order to resolve occlusion situations. At this point, we highlight the significant aspects of the proposed system in Chapter 3:

- Spatio-temporal Oriented Energy (SOE) features, representing the motion dynamics of the targets, are applied within a 'Bayesian model' to determine the visibility status of the targets in the course of tracking. Various modes of occlusion, namely the 'Partial' and 'Full' occlusion, are discriminated through this model.

- Perceiving the occlusion mode is very beneficial in video surveillance of public areas, due to frequent short-term occlusion events. The proposed 'Bayesian model' and the estimated occlusion modes are utilized as a means to establish an adaptive updating mechanism, in order to protect the target models against corruption in occlusion and drift situations.

- An integral Gaussian motion model is defined for every target, based on SOE features of the target pixels. The proposed model is used for estimating an occlusion mask which identifies the visible target pixels, against the occluded parts and the background pixels. Hence an improved tracking performance is expected, specially in occlusion situations, by concentrating the tracking optimization on visible parts of the targets.

- Due to invisibility of the targets in 'Full Occlusion' events, the tracking strategy is temporarily changed to a 'tracking by prediction' and concurrent searching for the lost target. In this situation, the system looks for the SOE energy blob of the disappeared target, for reporting reacquisition of the lost target.

- We demonstrate that maintaining a valid up-to-date template by protecting the target model against corruption or undesired changes in challenging real world scenarios, enables the system to perform more robustly and compete with state-of-the-art trackers.

- Utilizing motion dynamics of the targets, for the purpose of occlusion modeling, alleviates the confusion problem among similar targets in occlusion events.

Suggestions for future research

In this part we discuss some of the shortcomings of the proposed system, which may be addressed in future research and suggest potential extensions to the current work:

- One limitation of the proposed system, is the lack of an explicit appearance model in the occlusion modeling framework. The suggested occlusion model in the current system is solely based on motion dynamics. This model mitigates the confusion problem, when similar targets with different motion dynamics overlap. However for targets moving in a similar direction, an occlusion event may not be detected. In other words, the proposed occlusion model does not discriminate between the occluding targets with similar motion dynamics. In such scenarios, the

system is still vulnerable to template contamination and distraction in long term occlusions, although in short term the tracker might perform well due to the gradual template updating scheme. To overcome this limitation, we suggest to incorporate an efficient appearance model in the framework for the purpose of occlusion analysis. The appearance and motion models may play a complementary role for resolving occlusion situations, when the targets have similar motion or appearance.

- Another issue is about reacquisition of the targets following the full occlusion events, where the targets are totally invisible within a period of time. We proposed a tracking-by-prediction strategy when there is no visible target for tracking, while the system explores the surrounding area to recover the lost target. The current recovery scheme is based on exploring the target's motion blob corresponding to its motion model prior to obstruction. However incorporating an appearance model in the reacquisition scheme, may improve robustness of the system for the purpose of crowd analysis. A combination of motion and appearance or even automatic selection of the optimal mode for the target recovery after full occlusions, are the issues which are worth further investigation.

- The experimental evaluation of the proposed system has been conducted on surveillance videos from stationary cameras. The framework assumptions suggest that it may be applicable to moving camera videos, subject to smooth camera motion. In other words unstable or shaky videos may not provide suitable grounds for estimating motion models of the targets, unless video stabilization is applied prior to tracking process. video stabilization is supposed to eliminate the undesirable camera motions from the video, which are caused by hand-held or mechanical vibrations. Investigating the applicability of the proposed system to scenarios from moving cameras, is another potential extension to the current work.

6.2 Occlusion Handling in Detection-based Tracking

Chapter 4 introduced an occlusion handling method in the context of multiple target tracking by detection, based on a spatio-temporal clustering method. Such tracking approaches are composed of an object detector, followed by a data association method. In our system, a state-of-the-art pedestrian detector with an off-line trained full-body model, is used as the first stage of the system. Occlusions and confusions of interacting targets are resolved in the data association stage. In this part, we underscore the representative contributions of the proposed system in Chapter 4:

- The detection results from a pedestrian detector prior to NMS (Non-Max-Suppression), is processed through an efficient spatio-temporal clustering framework, as a substitute for NMS process. The spatial and temporal terms are combined within a principled framework, to improve NMS method in occlusion situations and simultaneously compensate for the general detection errors, such as missed/false detections. Temporal terms formulate motion and scale consistency of the tracked targets in successive frames. The proposed cost function, obtains a light-weight closed form solution which can be solved in real-time with a standard CPU.

- The proposed cost function is based on a bounded distance metric, which performs a local clustering within a neighborhood of each cluster center. The remaining non-associated members at each round of clustering optimization, is used for detecting new emergent targets and establishing new clusters. The non-associated members are processed with standard NMS and evaluated with a specific confidence metric, to increase the reliability of the instantiated clusters. The confidence metric is based on 'Depth/Height' probability of the NMS results and their 'Detection Frequency', which implies the number of suppressed detections per NMS result.

- The clustering results are post-processed within a consolidation framework, for resolving the remaining issues and consolidating the estimated tracks. The consolidation method uses the scene 'Entry/Exit'

information as a metric to determine the incomplete tracks, which require further processing. The incomplete tracks which do not start or end at Entry/Exit areas are integrated through joining consistent tracks. Furthermore a confidence score is defined for every track based on: **(i)** summation of the clusters' Depth/Height confidence along the track, which contains the track length information; and **(ii)** an 'Occlusion Matrix' denoting the overlap of clusters per frame. This confidence score is used for detecting spurious tracks and removing low confidence ones.

- Due to utilizing a full-body pedestrian model for detection, high resolution details are not required for good performance. Consequently the proposed framework is appropriate for existing infrastructures with low resolution cameras. Compatibility of the system with low resolution videos, along with its real-time performance, make the proposed framework suitable for practical video surveillance.

Suggestions for future research

In this section we discuss some of the imperfections of the proposed framework, which may be addressed in future research, and suggest potential extensions to the current system:

- Temporal terms of the clustering cost function, associate the motion and scale of the estimated clusters in consecutive frames. In normal situation where there is no interaction among the targets or in occlusion events where the interacting targets have different scales or motion directions, the proposed temporal terms are capable for resolving the ambiguity. However in cases that the overlapping targets have similar scales and motion directions, the proposed temporal terms do not suffice to discriminate among the occluding targets. Hence identity switches among the targets are probable in such occasions. Incorporating a proper appearance model within the temporal terms of the clustering cost function, may resolve the remaining confusions in such events. In other words, an appearance model may play a complementary role alongside the other temporal terms in resolving the ambiguities, when the overlapping targets have similar scales and motion directions.

- In the current consolidation framework, immature tracks are completed by joining to other consistent tracks. Consistency of tracks is identified based on the motion and scale coherency of their corresponding clusters at the connection area. However, when more than two targets with similar motion directions overlap, an appearance model may help to resolve the confusion of tracks. Hence we suggest to incorporate an appearance model for detection of the consistent tracks, in order to improve the robustness of the system for the purpose of crowd analysis.

- The current consolidation scheme takes advantage of the Entry/Exit information, for detection of the incomplete tracks. In this work, Entry/Exit areas are manually defined as a one-time set-up through calibration of cameras. A possible extension to this approach, is to include an online incremental learning scheme for gradual grasping of scene geometry and Entry/Exit information, through the flow of crowd in time.

- Information of the scene geometry such as the Entry/Exit areas, can be very helpful in video surveillance scenarios with static cameras. However, for extending the framework to moving cameras, such as in working robots, we need to develop other robust schemes for track consolidation. we suggest to use a track confidence score composed of: **(i)** Depth/Height confidence along the track; **(ii)** Consistency of appearance along the track; and **(iii)** Overlap of clusters along the track.

- The proposed occlusion handling framework is designed for sparse crowds, where the targets are expected to be completely visible for a period of time among their presence in the scene. This provides enough basis for instantiation of new clusters, identifying their correct scale and motion pattern and for capturing a valid appearance model of the targets. To improve the performance in medium and dense crowds, it may be beneficial to apply flow information like optical flow, SOE features or other similar cues, for identifying the coherently moving targets and to estimate their correct motion patterns.

- This work was a first step towards solving the more challenging problem of real time crowd analysis. It is worth to investigate how far

the above mentioned generalizations can improve the performance in medium and dense crowds.

6.3 GPU-based Hardware Acceleration

Chapter 5 presented a high throughput GPU engine for accelerating the computation and extraction of 2D multi-channel Gabor features. Evaluation of the system indicates a real-time performance for a 12 channel non-separable filter on VGA size frames. The speedup gain demonstrated by the GPU engine, suggests that the separable multi-channel SOE features used in preceding chapters, may be also extracted in real-time. The main contribution in the proposed hardware accelerator are:

- The proposed engine provides an adaptive kernel, supporting non-unit aspect ratios or non-square kernels. More clearly, the engine configures the kernel size per channel, for efficient computation and accurate estimation of the feature channels.
- The parallel structure of the GPU kernel, is similar to initial layers of the visual cortex, composed of 'Simple' and 'Complex' cells. Simple units perform filtering operation in separate channels implying different scales and orientations. Complex units comprise the final layer of the processor, which aggregates the features provided by Simple units, at common orientations and scales. Aggregation is performed through feature pooling with 'Maximum' and 'Histogram' methods, that are widely applied in various applications. The abstract information provided by the Complex layer of the engine, describe the orientation structure of the input image with some level of local invariance.

Suggestions for future research

We suggest to use application specific metrics for investigating the effects of filter size variations, elliptic Gabor functions and non-square filter masks, on the performance of different tasks. Then the optimum kernel figures for the specific problems may be introduced, to efficiently use the flexibility of the kernel towards achieving the best performance for real time applications.

Bibliography

[AB85] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America - Optics Image Science and Vision*, 2(2):284–299, 1985.

[AJP00] L. H. A. Jain, S. Prabhakar and S. Pankanti. Filterbank-based fingerprint matching. *IEEE Transactions on Image Processing*, 9(5):846859, 2000.

[ARS06] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 798–805, 2006.

[ARS08] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[AS11] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1272, 2011.

[ATB09] G. Amayeh, A. Tavakkoli, and G. Bebis. Accurate and efficient computation of gabor features in real-time applications. In *Proceedings of Advances in Visual Computing, Lecture Notes in Computer Science*, volume 5875, pages 243–252, 2009.

[Avi01] S. Avidan. Support vector tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 184–191, 2001.

[Avi04] S. Avidan. Support vector tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(8):1064–1072, 2004.

[Avi05] S. Avidan. Ensemble tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 494–501, 2005.

[Avi07] S. Avidan. Ensemble tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):261–271, 2007.

[AWST05a] V. Areekul, U. Watchareeruetai, K. Suppasriwasuseth, and S. Tantaratana. Separable gabor filter realization for fast fingerprint enhancement. In *IEEE Intl. Conference On Image Processing*, volume 3, page 253256, 2005.

[AWST05b] V. Areekul, U. Watchareeruetai, K. Suppasriwasuseth, and S. Tantaratana. Separable gabor filter realization for fast fingerprint enhancement. In *International Conference on Image Processing - ICIP*, pages 253–256, 2005.

[AWT04] V. Areekul, U. Watchareeruetai, and S. Tantaratana. Fast separable gabor filter for fingerprint enhancement. In *Intl. Conf. on Biometric Authentication*, page 403409, 2004.

[BA83] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[BA96] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[BAHH92] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *European Conference on Computer Vision ECCV*, pages 237–252, 1992.

[Bat04] J. P. Batista. Tracking pedestrians under occlusion using multiple cameras. In *Image Analysis and Recognition*, pages 552–562. Springer Berlin Heidelberg, 2004.

[BB95] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Comput. Surv.*, 27(3):433–467, 1995.

[BFP14] L. M. G. Brown, R. S. Feris, and S. Pankanti. Temporal non-maximum suppression for pedestrian detection using self-calibration. In *International Conference on Pattern Recognition, ICPR*, pages 2239–2244, 2014.

[BHPD13] X. P. Burgos-Artizzu, D. Hall, P. Perona, and P. Dollár. Merging pose estimates across space and time. In *British Machine Vision Conference, BMVC*, 2013.

[BJ98] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[BK96] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions? In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 270–277, 1996.

[BL12] T. Bai and Y. Li. Robust visual tracking with structured sparse representation appearance model. *Pattern Recognition*, 45(6):2390–2404, 2012.

[BS08] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal of Image and Video Processing*, 2008, 2008.

[Bur81] P. J. Burt. Fast filter transforms for image processing. *Computer Graphics and Image Processing*, 16(1):20–51, 1981.

[BWLJ12] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1830–1837, 2012.

[BYB11] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, 2011.

[Can08] K. J. Cannons. A review of visual tracking. Technical report, YORK University, Department of Computer Science and Engineering, 2008.

[Can10] K. J. Cannons. Supplemental video results, pixelwise soe tracker eccv2010. http://www.cse.yorku.ca/vision/research/oriented-energy-tracking/, 2010.

[CAV04] Caviar test case scenarios, ec funded caviar project/ist 2001. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/, 2004.

[CGO00] T. Chang, S. Gong, and E. Ong. Tracking multiple people under occlusion using multiple cameras. In *Proceedings of the British Machine Vision Conference, BMVC*, pages 1–10, 2000.

[CGW10] K. J. Cannons, J. M. Gryn, and R. P. Wildes. Visual tracking using a pixelwise spatiotemporal oriented energy representation. In *Lecture Notes in Computer Science (ECCV)*, volume 6314, pages 511–524, 2010.

[CL03] R. T. Collins and Y. Liu. On-line selection of discriminative tracking features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 346–352, 2003.

[CPB09] P. Chockalingam, S. Pradeep, and S. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *IEEE International Conference on Computer Vision*, pages 1530–1537, 2009.

[CRM00] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 142–149, 2000.

[CRM03] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[CS10] D. M. Chu and A. W. M. Smeulders. Thirteen hard cases in visual tracking. In *IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS, Boston, MA, USA*, pages 103–110, 2010.

[CSD+08] V. Cevher, A. C. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *European Conference on Computer Vision ECCV*, pages 155–168, 2008.

[CUD09] Nvidia cuda programming guide. Technical Report VER 2.3.1, NVIDIA, 8 2009.

[CW07] K. J. Cannons and R. P. Wildes. Spatiotemporal oriented energy features for visual tracking. In *Lecture Notes in Computer Science (ACCV)*, volume 4843, pages 532–543, 2007.

[CW14] K. J. Cannons and R. P. Wildes. The applicability of spatiotemporal oriented energy features to region tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):784–796, 2014.

[DABP14] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014.

[DG05] K. G. Derpanis and J. M. Gryn. Three-dimensional nth derivative of gaussian separable steerable filters. In *IEEE International Conference on Image Processing, ICIP*, pages 553–456, 2005.

[Dia06] J. Diaz. *Multimodal bio-inspired vision system, High performance motion and stereo processing architecture*. PhD thesis, 2006.

[DLDW12] K. G. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *IEEE Conference*

on Computer Vision and Pattern Recognition, pages 1306–1313, 2012.

[DSCW10] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1990–1997, 2010.

[DT00] S. L. Dockstader and A. M. Tekalp. Tracking multiple objects in the presence of articulated and occluded motion. In *Workshop on Human Motion*, page 88, 2000.

[DT01] S. L. Dockstader and A. M. Tekalp. Multi-view spatial integration and tracking with bayesian networks. In *ICIP*, pages 630–633, 2001.

[DT05] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 886–893, 2005.

[DV05] M. Do and M. Vetterli. The contourlet transform: An efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2107–2116, 2005.

[DW09] K. G. Derpanis and R. P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–239, 2009.

[DW10] K. G. Derpanis and R. P. Wildes. Dynamic texture recognition based on distributions of spacetime oriented structure. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 191–198, 2010.

[DW11] K. G. Derpanis and R. P. Wildes. Classification of traffic video based on a spatiotemporal orientation analysis. In *IEEE Workshop on Applications of Computer Vision*, pages 606–613, 2011.

[DW12] K. G. Derpanis and R. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(6):1193–1205, 2012.

[DWSP09] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *IEEE Computer Society Conference on Computer Vision and Pattern*, pages 304–311, 2009.

[DWSP12] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, 2012.

[EESG10] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, pages 990–997, 2010.

[EJCY06] D. L. D. E. J. Candes, L. Demanet and L. Ying. Fast discrete curvelet transforms. *Multi-scale Modeling and Simulation*, 5(3):861899, 2006.

[FA91] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991.

[FGMR10] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[FP81] M. Fahle and T. Poggio. Visual hyperacuity: spatiotemporal interpolation in human vision. *Proceedings of the Royal Society of London-B*, 213(1193):451–477, 1981.

[Gab46] D. Gabor. Theory of communication. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[GGB06] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *Proceedings of the British Machine Vision Conference*, pages 47–56, 2006.

[GPK11] T. Gao, B. Packer, and D. Koller. A segmentation-aware object detection model with occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1361–1368, 2011.

[GSRL98] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 22–29, 1998.

[HB96] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 403–410, 1996.

[Hee88] D. Heeger. Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, 1(4):279–302, 1988.

[HJZ+11] Z. Han, J. Jiao, B. Zhang, Q. Ye, and J. Liu. Visual object tracking via sample-based adaptive sparse representation (adasr). *Pattern Recognition*, 44(9):2170–2183, 2011.

[HX05] Y. HUANG and M. Xie. A novel character-recognition method based on gabor transform. In *International Conference on Communications, Circuits and Systems*, volume 2, pages 815–819, 2005.

[iLI06] i-LIDS, UK government benchmark datasets for automated surveillance. http://www.homeoffice.gov.uk/publications/science/cast/ilids-brochure/, 2006.

[JF91] A. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern Recognition*, 24(12):11671186, 1991.

[JLY12] X. Jia, H. Lu, and M.-H. Yang. Visual tracking via adaptive structural local sparse appearance model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1822–1829, 2012.

[Kal60] R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering*, 82-D:32–45, 1960.

[Kin99] N. Kingsbury. Image processing with complex wavelets. *Philosophical Transactions A, Royal Society of London*, 357(1760):2543–2560, 1999.

[KL11] J. Kwon and K. M. Lee. Tracking by sampling trackers. In *IEEE International Conference on Computer Vision*, pages 1195–1202, 2011.

[KMM12] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2012.

[KNHH11] S. Kwak, W. Nam, B. Han, and J. H. Han. Learning occlusion with likelihoods for visual tracking. In *IEEE International Conference on Computer Vision*, pages 1551–1558, 2011.

[Kub95] T. Kubota. *Orientational Filters For RealTime Computer Vision Problems*. PhD thesis, 1995.

[Lin94] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

[LK81] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981.

[LW02] C. LIU and H. WECHSLER. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002.

[LW03] C. Liu and H. Wechsler. Independent component analysis of gabor features for face recognition. *IEEE Transactions on Neural Networks*, 14(4):919928, 2003.

[LYY13] X. Lu, Y. Yuan, and P. Yan. Robust visual tracking with discriminative sparse learning. *Pattern Recognition*, 46(7):1762 – 1771, 2013.

[Mar80] S. Marcelja. Mathematical description of the responses of simple cortical cells. *Optical Society of America*, 70:1297–1300, 1980.

[MBM+95] D. Murray, K. Bradshaw, P. McLauchlan, I. Reid, and P. Sharkey. Driving saccade to pursuit using image motion. *International Journal of Computer Vision*, 16(3):205–228, 1995.

[MBTG13] M. Mathias, R. Benenson, R. Timofte, and L. J. V. Gool. Handling occlusions with franken-classifiers. In *IEEE International Conference on Computer Vision*, pages 1505–1512, 2013.

[ML09] X. Mei and H. Ling. Robust visual tracking using l1 minimization. In *IEEE International Conference on Computer Vision*, pages 1436–1443, 2009.

[ML11] X. Mei and H. Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2259–2272, 2011.

[MLW+11] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai. Minimum error bounded efficient l1 tracker with occlusion detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1257–1264, 2011.

[MRS14] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(1):58–72, 2014.

[MWT02] L. Ma, Y. Wang, and T. Tan. Iris recognition based on multichannel gabor filtering. In *IEEE Asian Conference on Computer Vision*, page 279283, 2002.

[NC10] T. T. Nguyen and H. Chauris. Uniform discrete curvelet transform. *IEEE Transactions on Signal Processing*, 58(7):3618–3634, 2010.

[NNPT98] O. Nestares, R. Navarro, J. Portilla, and A. Tabernero. Efficient spatial-domain implementation of a multiscale image representation based on gabor functions. *Transaction of Electronic Imaging*, 7(1):166173, 1998.

[NO06] T. Nguyen and S. Oraintara. A shift-invariant multiscale multidirection image decomposition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 153–156, 2006.

[NO08] T. T. Nguyen and S. Oraintara. The shiftable complex directional pyramid, part 1: The theoretical aspects. *IEEE Transactions on Signal Processing*, 56(10):46514660, 2008.

[NSHY08] S. Nejhum Shahed, J. Ho, and M. Yang. Visual tracking with histograms and articulating blocks. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2008.

[Ope09] Opencl programming guide for the cuda architecture. Technical Report VER 2.3, NVIDIA, 8 2009.

[PBE+06] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Toward Category-Level Object Recognition*, pages 29–48, 2006.

[PC07] M. Pham and T. Cham. Online learning asymmetric boosted classifiers for object detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[PET07] Pets 2007 benchmark data. http://www.cvg.rdg.ac.uk/PETS2007/data.html, 2007.

[PHC08] M. Pham, V. D. Hoang, and T. Cham. Detection with multi-exit asymmetric boosting. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[PK97] N. Petkov and P. Kruizinga. Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells. *Biological Cybernetics*, 76(2):83–96, 1997.

[RC13] S. Rujikietgumjorn and R. T. Collins. Optimized pedestrian detection for multiple and occluded people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3690–3697, 2013.

[RLLY08] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008.

[RS91] K. Rangarajan and M. Shah. Establishing motion correspondence. *CVGIP:Image Understanding*, 54(1):56–73, 1991.

[Sal14] A. Saleh. Partially occluded pedestrian classification using histogram of oriented gradients and local weighted linear kernel support vector machine. *IET COMPUTER VISION*, 8(6):620–628, 2014.

[SBK05] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6):123–151, 2005.

[SFC+11] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition-CVPR*, pages 1297–1304, 2011.

[SHT+06] A. W. Senior, A. Hampapur, Y.-l. Tian, L. M. Brown, S. Pankanti, and R. M. Bolle. Appearance models for occlusion handling. *Image and Vision Computing*, 24(11):1233–1243, 2006.

[Sim93] E. Simoncelli. *Distributed representation and analysis of visual motion*. PhD thesis, 1993.

[Sit64] R. W. Sittler. An optimal data association problem in surveillance theory. *IEEE Transactions on Military Electronics*, 8:125–139, 1964.

[SJ87] I. Sethi and R. Jain. Finding trajectories of feature points in monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):56–73, 1987.

[SS71] R. Singer and J. Stein. An optimal tracking filter for processing sensor data of imprecisely determined origin in surveillance systems. In *IEEE Conference on Decision and Control*, pages 171–175, 1971.

[SS90] V. Salari and I. K. Sethi. Feature point correspondence in the presence of occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):87–91, 1990.

[SS03] K. Shafique and M. Shah. A non-iterative greedy algorithm for multi-frame point correspondence. In *IEEE International Conference on Computer Vision*, pages 110–115, 2003.

[ST94] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 593–600, 1994.

[SWP05] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 994–1000, 2005.

[TAS12] S. Tang, M. Andriluka, and B. Schiele. Detection and tracking of occluded people. In *British Machine Vision Conference*, pages 1–11, 2012.

[TK91] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University Technical Report, April 1991.

[TSK00] H. Tao, H. Sawhney, and R. Kumar. Dynamic layer representation with applications to tracking. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 134–141, 2000.

[VO08] A. P. N. VO. *COMPLEX DIRECTIONAL WAVELET TRANSFORMS: REPRESENTATION, STATISTICAL MODELING AND APPLICATIONS*. PhD thesis, 2008.

[WC05] Y. Wang and C. Chua. Face recognition from 2d and 3d images using 3d gabor filters. *Image and Vision Computing*, 23(11):1018–1028, 2005.

[WCXY12] Q. Wang, F. Chen, W. Xu, and M.-H. Yang. Online discriminative object tracking with local sparse representation. In *IEEE Workshop on Applications of Computer Vision*, pages 425–432, 2012.

[WHY09] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *IEEE 12th International Conference on Computer Vision*, pages 32–39, 2009.

[WLYY11] S. Wang, H. Lu, F. Yang, and M.-H. Yang. Superpixel tracking. In *IEEE International Conference on Computer Vision*, pages 1323–1330, 2011.

[WMSS10] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1030–1037, 2010.

[WRC08] K. Woodbeck, G. Roth, and H. Chen. Visual cortex on the gpu: Biologically inspired classifier and feature descriptor for rapid recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[WTSB12] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke. Coupling detection and data association for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1948–1955, 2012.

[WU08] B. WU. *Part based Object Detection, Segmentation, and Tracking by Boosting Simple Feature based Weak Classifiers*. PhD thesis, 2008.

[WWRS11] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1993–2000, 2011.

[WYG+09] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transaction of Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[YJS06] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4), 2006.

[YL01] R. Young and R. Lesperance. The gaussian derivative model for spatial-temporal vision: I. cortical model. *Transaction of Spatial Vision*, 14(3-4):261–319, 2001.

[YOR10] Tracking dataset of york vision lab. http://www.cse.yorku.ca/vision/research/visual-tracking/, 2010.

[You86] R. Young. Simulation of human retinal function with the gaussian derivative model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 564–569, 1986.

[ZYSL13] S. Zhang, H. Yao, X. Sun, and X. Lu. Sparse coding based visual tracking: Review and experimental comparison. *Pattern Recognition*, 46(7):1772 – 1788, 2013.