CROWD DENSITY ESTIMATION AND ANALYSIS

Ph.D. Thesis

SONU LAMBA

(ID No. 2015RCP9003)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR

December, 2019

Crowd Density Estimation and Analysis

submitted in

fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Sonu Lamba

ID: 2015RCP9003

Under the Supervision of

Dr. Neeta Nain

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY JAIPUR

December, 2019

DECLARATION

I, **Sonu Lamba**, declare that this thesis titled,**"CROWD DENSITY ESTIMATION AND ANALYSIS"** and the work presented in it, are my own, I confirm that:

- This work was done wholly or mainly while in candidature for a research degree t this university.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always give. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself, jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date:

Sonu Lamba
(2015RCP9003)

CERTIFICATE

This is to certify that the thesis entitled, **"CROWD DENSITY ESTIMATION AND ANALYSIS"** being submitted by **Sonu Lamba (2015RCP9003)** is a bonafide research work carried out under our supervision and guidance in fulfillment of the requirement for the award of the degree of **Doctor of Philosophy** in the Department of Computer Science & Engineering, Malaviya National Institute of Technology Jaipur, India. The matter embodied in this thesis is original and has not been submitted to any other university or institute for the award of any other degree.

Place: Jaipur
Date:

Dr. Neeta Nain
Associate Professor
Computer Science and Engineering
MNIT Jaipur

ACKNOWLEDGMENT

It gives me immense pleasure to express gratitude and regards to all those people who supported me during the course of this doctoral research work at MNIT Jaipur. I acknowledge the involvement and contribution of each one of them. I would first like to thank God who gave me the grace and privilege to pursue this program and successfully complete it in spite of many challenges faced. I express my heart-felt gratitude to my supervisor Dr. Neeta Nain for her constant inspiration, encouragement, valuable advices, enormous support and blessings. It is needless to say without her, this research work would have not been possible. The best time and fruitful discussion with her have immense contributions in the process of completion of my Ph.D work on time that would be treasured throughout my life.

My special thanks to the members of Doctoral Research Ethics Committee (DREC), Dr. Neeta Nain, Dr. Dinesh Gopalani, Dr. Yogesh Kumar Meena, Dr. Santosh Viparthi for their constructive criticisms and valuable suggestions. I am also grateful to all other faculty members of the department for their inspirational and ideas at various stages. I am also thankful to the officials of the department for their kind help in official work.

Let me express special thanks to Head of the Department, Dr. Pilli Emmanuel Shubhakar, for the keen support and consistent encouragement in our academic activities. I am extremely thankful to Prof. Udaykumar R Yaragatti, Director, MNIT Jaipur for providing me infrastructural facilities to work in, without which this work would not have been possible.

I would also like to thank all the referees who reviewed this work as pieces of it were submitted to international journals and conferences. Their detailed reviews, constructive criticism and excellent advice have improved both the presentation and content of this thesis.

I would like to thank my co-scholars Shweta, Tapas, Riti, Tanvi, Anamika, Maninder and other research colleagues of the department for their loving cooperation, positive criticism, excellent advice, consistent support and consideration during the preparation of this thesis. I express my gratitude to my estimable friend cum sister Shweta Saharan for her perpetual support, motivation, love, care and to spend quality time together, also having cheered me up always.

I feel a deep sense of gratitude for my parents Mr. Vidhan Chandra Lamba and Mrs. Sumitra Singh, my brothers Rahul and Deepak and my uncle and aunt Mr.

Rajpal Lamba & Mrs. Subhita Devi who formed part of my vision and taught me good things that really matters in life. I thank my father for his unconditional love and support throughout my educational career. Their patience and sacrifice will remain my inspiration throughout my life. And most of all, I owe my deepest gratitude to my husband, Mr. Sandeep Dhinwa, for his affection, encouragement, understanding and patience. He supported me without any complaint or regret that enabled me to complete my Ph. D thesis. My thesis acknowledgement would be incomplete without thanking my baby-girl, Saanvi, whose smiling face always made me happy and inspired me. Having her midway during my Ph.D. was certainly not easy for me but she has made my life wonderful. I offer my regards to my father in-law Mr. Guljharee Lal and mother in-law Mrs. Geeta Devi for their constant support to complete this degree. Finally, I gratefully, acknowledge one and all who are directly or indirectly involved to shape this research work.

Abstract

With the exponential growth in worldwide population, crowd analysis has become a dire need for public safety and security perspectives. Visual analysis of dense crowd is particularly difficult because of the high density of individuals, severe occlusions, cluttered background, and fewer pixels per individuals which hardly exist in regular surveillance environments.

This thesis intends to tackle these issues both in images and videos of intensely dense crowds. The dense crowd consists of hundreds to thousands of people per image or scene. The target is to handle the elementary problems of crowd detection, density estimation, flow segmentation and anomaly detection in such images and videos using texture features which are extracted from the crowded scenes.

In this dissertation, the study of crowd analysis includes crowd density estimation and crowd scene understanding. We also find out the limitations of approaches present in existing literature to bring out the solution of fully automatic crowd density estimation and analysis of crowd scenes. This dissertation focuses on four topics, namely crowd detection, crowd density estimation, flow segmentation, and anomalous scene detection.

For density estimation, we present to leverage the clues obtained from multiple sources, to figure out the density of people existent in exceptionally dense crowd images. Our approach depends on multiple sources such as head detection with low confidence, recurrence of texture elements by using frequency domain, wavelet and scale invariant feature transform (SIFT) descriptor to measure the density count. The information obtained from different sources trains a support vector machine (SVM), which generates a patch count estimation. Furthermore, we apply Markov Random Field (MRF) on image patches to obtain uniformity in counts in local neighborhoods and across scales. We tested this approach on four different datasets such as Shanghai Tech_A, UCF _CC_50, extended UCF_CC_100 and UCSD.

The former three datasets are a crisp contrast to existing crowd datasets used in literature which contains almost hundreds or tens of individuals in crowd images. The latter UCSD dataset is used to test the robustness of our technique in the low-density crowd too. We also compare our method with both traditional and convolutional neural network (CNN) based approaches. Low computational complexity indicates that our technique provides decent performance rate and can be employed in real-world applications.

Our experimental results validate the adequacy and efficiency of the intended methodology by measuring the density of people in images of high-density crowds in contrast to existing methods which are merely suitable for low to medium crowd density.

Understanding crowd scene is another important task in crowd analysis where the goal is to segment dominant flow pattern for preventing accidents as well as anomalous scene detection for implementing evacuation plans essential in case of over-crowdedness, clogging or riots in the urban regions.

In crowd flow segmentation, the objective of this work is to segment crowd flow while simultaneously considering the newly appearing object in the scene. It is accomplished by applying an updating procedure after a temporal window of a certain number of frames. Furthermore, the essential aspect of the presented method is taking advantages of foreground segmentation by active contouring scheme. From the segmented region, trajectories are extracted by considering block level tracking instead of individual point tracking.

The presented approach is highly successful in high-density crowd scenes due to block level tracking, which is a primary reason for decreasing the performance of standard tracking by individual detection approaches. We cluster the extracted trajectories by designing a novel clustering algorithm, especially applicable to a high-density crowd scene. We extract the shape, location, direction and density features of trajectory patterns. Next, each pixel is labeled depending upon their motion pattern to segment the final crowd flow and eventually, the crowd flow segments are analyzed to understand the flow. As the approach does not demand high computational resources, it can be used in segmenting flow patterns in real time crowd video sequences.

Once crowd flow patterns are segmented, we then use the extracted tracklets for detecting anomalous scene in dense crowd videos. The purpose of crowd scene analysis is early detection or prediction of anomalous events which could lead to potentially dangerous situations that threaten the safety of individuals. By early detection of anomalous scenes, potentially threatening reactions can be reduced or prevented. This work considers the anomalous crowd scene detection by performing a statistical analysis on the oriented tracklets of the crowd. We adopt two statistical parameters as entropy and temporal occupancy for anomaly detection. Our idea is to model a normal crowd scene based on these parameters and any discrepancy in the normal crowd behavior is detected as anomalous. Our method is evaluated on various crowd video sequences containing medium to high-

density and outperforms on existing approaches without prior learning of crowd flow patterns.

We validate the reliability of our algorithms by evaluating the performance of crowd detection, density estimation, as well as flow segmentation and anomalous scene detection on different datasets comprising thousands of persons in challenging crowd scenarios.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Due to experiences of exponential expansion in the worldwide population, crowd scenes have been more common than ever. Crowd scene comprises a significantly large number of individuals assembled collectively in one place. Such places often acquire mass gathering events such as political speeches, public demonstrations, music concerts, marathons, and religious gatherings. There are intrinsic dangers coupled with each large public gathering. Every year there are stories of overcrowding and crushing incidents from around the world. Each year individuals are injured and die in crowd-related mishaps, even in some cases at planned occasions. Parsing through a timeline of crowd-related disasters over the past decade paints a sad picture. For instance, some well-known examples of crowd calamities include Chamunda Devi temple stampede at Rajasthan (India) in 2008, Love Parade disaster at Germany in 2010, Khmer water festival-Penh in Cambodia (2010), Boston marathon bombing in 2013, Mina-Mecca stampede in 2015 and Mumbai railway station stampede in 2017 are illustrated in Figure 1.1. A few more illustrative examples of worldwide deadliest stampedes are shown in Table 1.1 where the number of deaths is reported as high as 2262 persons.

To prevent these unfortunate events or tragedies from happening, crowd phenomenon is becoming an important topic for research. The learning of the crowd has therefore caught the attention of multidisciplinary research from civil, computer science, physics, psychology, and biology. In computer science, computer vision researchers have gained much attention towards an automatic analysis of crowd in the past decade.

Crowd analysis has numerous real-world applications in public space construction, virtual simulation, visual surveillance, and especially in crowd management. To

Figure 1.1: Some well-known cases of crowd related tragedies at venue of mass gatherings: first row left to right indicates Chamunda Devi temple stampede, Rajasthan-India (2008), Love Parade disaster-Duisburg, Germany (2010) and Khmer water festival-Penh, Cambodia (2010). The second row left to right indicates Boston marathon bombing, Massachusetts-United States (2013), Mina-Mecca, Saudi Arabia (2015) and Mumbai railway station stampede(2017) respectively.

observe a big area of the crowd, various surveillance cameras are set up around the world that generate a large number of video streams. Manual analysis of these massive amounts of video data and to make intelligent decisions for emergencies are impossible for any security personnel. Even, a human observer may fail to notice a suspicious event. Therefore, there must be an automated way to analyze the data generated by surveillance systems.

The visual analysis of crowd incorporates numerous computer vision tasks such as object detection, counting (density estimation), tracking, and understanding the behavior of various objects or crowd scenes, as shown in Figure 1.2. To perform these tasks, many vision-based algorithms are developed, but they were primarily designed for frequent scenes with a low density of crowd [Hu et al. (2006, 2004)]. The complexity of computer vision tasks augments disproportionately relying on the number of persons forming the crowd. For low-density crowd scenes, detection and tracking methods might work very well due to the apparent visibility of people. However, detection and tracking fail when it comes to high-density crowded scenes with hundreds to thousands of individuals, and also increases computational complexity. Therefore, the applicability of a particular computer vision technique relies upon the density and structure of the crowd.

Dense crowds put forward a set of challenges in visual analysis like cluttered background, fewer pixels per person, severe occlusions and perspective effects, etc.

Table 1.1: List of notable crowd stampedes and crushes by death toll

S.No.	Year	Location	Deaths	Event
1	1883	Victoria Hall, Britain	180	Entertainment
2	1913	Italian Hall in Calumet City, Michigan	73	Religious
3	1964	National stadium in Lima, Peru	300	Sports
4	1989	Hillsborough Stadium in Sheffield	93	Sports
5	1990, 2004, 2015	Hajj pilgrimage in Mina, Mecca, Saudi Arabia	1426, 244, 2262	Religious
6	2003	New Jersey's Station Nightclub	96	Night Club
7	2005	Bridge stampede, Bagdhad	1000	Religious
8	2006	The Ultra arena in Manila, Philippines	73	Sports
9	2008	The Chamunda Devi Temple at top hill in Jodhpur, India	200	Religious
10	2013	Hindu Navratri festival, Datia, Madhya Pradesh, India	115	Religious
11	2016	Annual thanksgiving festival of the Oromo people in Ethiopia	300	Protest
12	2018	The El Paraso Social Club, Caracas, Venezuela	21	Protest

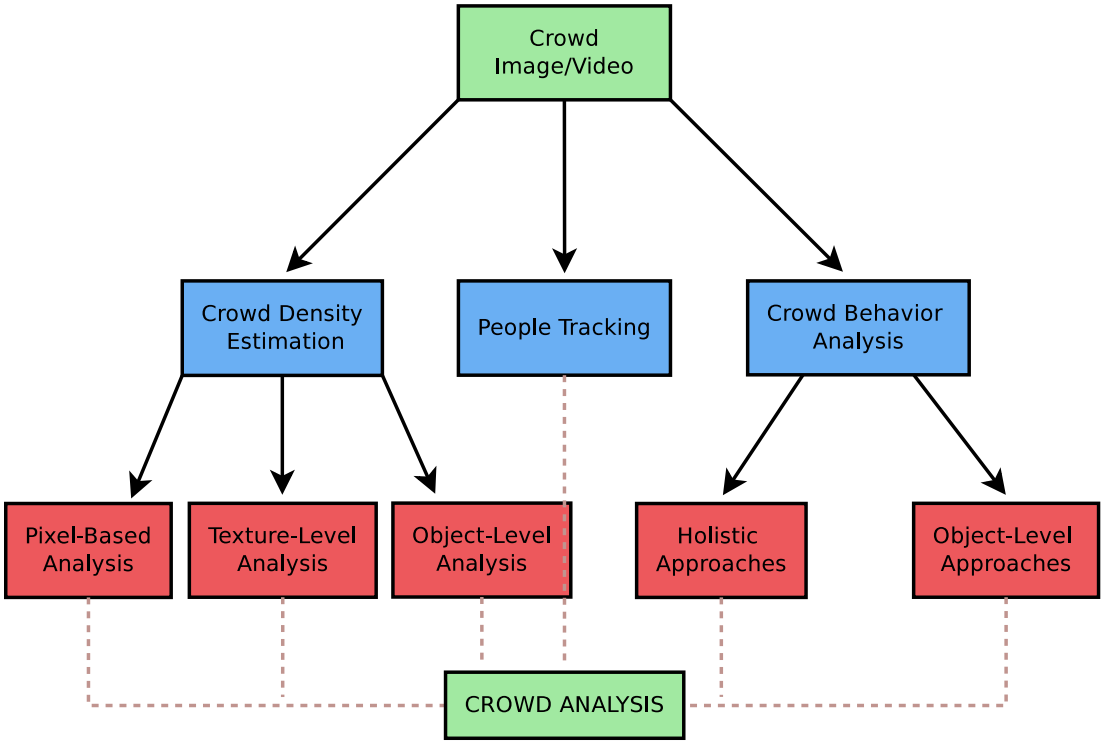

Figure 1.2: Schematic representation of the topics tackled in crowd analysis.

Therefore, in the dense crowd, the entire understanding of the scene or image is necessary. In this thesis, the study of crowd analysis includes crowd density estimation and crowd scene understanding. For density estimation, we explore the use of texture features in the dense crowd. Since it is observed that the closely coupled crowds of persons can be considered as a texture, and it appears irregular when viewed at a coarse-scale but with greater details of the texture, it follows a harmonic or regular pattern. We adopted holistic approaches for crowd scene understanding in terms of finding anomalous scenes and segmentation of crowd flow or motion patterns.

1.1 Crowd Density Estimation

Crowd density is the number of people within a unit area, such as people per square meter. Crowd density estimation is a prerequisite for all the measures of crowd analysis. Density is one of the primitive explanation of crowd status and, has applications in public space design inclusive of their expansion and amendments, by inspecting the counts of commuters that generally travel through these areas. It estimates the capacity of public space design and marks the tendency of abnormal changes in density over time. Manual counting of people in a dense crowd is a tremendously laborious task. The computer vision researchers have proposed various automated and semi-automated solutions for crowd density estimation.

Traditional crowd counting methods, [Chan et al. (2008); Ryan et al. (2009); Chan and Vasconcelos (2012)] estimate crowd density by individual pedestrian detection. Human detection is a basic function for density estimation, people tracking, and strange activity recognition. Reliable human detection is a key component of robust human tracking. In mass gatherings, the human body may be partially or completely occluded, which degrades the resultant performance. A face is the most evident part of the body which gets captured in the images since cameras are fixed at high altitude for improved surveillance. We developed a crowd face detection method that incorporates a skin color model, histogram of oriented gradient feature with Support Vector Machine (SVM). In this approach, the non-skin color objects are automatically discarded by the skin color model. SVM classifies the leftover objects based on the histogram of gradient oriented features, which simplify an object in such a way that the object generates nearly the same features when envisioned in diverse conditions. We find excellent results for the low-density crowd where faces are visible. However, the performance of our approach deteriorates

with an increase in crowd density or invisibility of frontal faces.

In the extremely dense crowded scenes, where each individual occupies very few pixels with severe occlusion, face, head, and full body detectors fail to detect the individuals. A single feature or detection method alone is not suitable to produce a correct crowd count due to small target size, low resolution, harsh occlusion between individuals, cluttered background, and perspective effects. Even the state-of-the-art like human, head, or face detectors perform poorly in such scenarios. High-density crowd images occupy very few pixels per target so that neither it can be detected nor its presence be verified in the location. These are the key requirement of the existing literature. The practicability of most existing techniques is bounded by the following two shortcomings:

1. Incapability to deal with crowds of hundreds or thousands of people.
2. Dependence on a single feature or detection method in the high-density crowd.

Therefore, to deal with a high-density crowd, we present a regression-based multisource approach that extracts effective texture features from crowd images to estimate the density of the crowd. We examined that the tightly coupled crowds of individuals can be considered as a texture, and the texture looks irregular when viewed at a coarse scale but as we go into finer details of the texture, it seems to follow a particular harmonic or regular pattern. The multi-source approach includes scale invariant feature transform (SIFT) descriptor, Fourier analysis, Wavelet analysis, and head detection. Initially, Fourier analysis is employed along with head detections, followed by Wavelet analysis and interest point based SIFT counts.

All these techniques are sequentially applied in local neighborhoods at multiple scales for avoiding the issues arising from irregularity in the observed textures arising from dense crowd images. The density count is performed at patch level to overcome the problems arising due to foreshortening and local geometric distortions. The estimated counts of each patch are accumulated while keeping the consistency constraints over the whole image. We gain the advantage of multiple counts from different approaches by analyzing the individual source. We obtain a confidence score from each source, and a Support Vector Regressor (SVR) is trained by using the ground truth annotations and computed features. Further, all the information sources are combined with their corresponding confidences, the count is independently computed at localized patches, and then the entire image

count is estimated with global consistency constraint. Since the counts are computed separately at different scales, the uniformity constraint is suitable for spatial neighborhoods along with direct neighbors at diverse scales.

We present a solution to gain uniformity at all scales to the whole image by applying multi-scale grid Markov Random Field (MRF) [Li (2009)] structure that enforces the count consistency constraint. The total crowd density count in an image is achieved by aggregating the patch counts of corresponding image patches in the grid structure. Our approach is adequate to estimate the crowd density in high-density crowd images. Several other tasks aimed at automated analysis of dense crowds require crowd scene analysis where the goal is to segment dominant flow patterns to prevent accidents, as well as anomalous scene detection from implementing evacuation plans essential in the unlikely event of over-crowdedness, clogging or in the case of riots in urban areas.

1.2 Crowd Scene Analysis

Automated analysis of crowd scenes has a significant role in ensuring public safety and better management of incidents comprising a huge crowd. It has various utilizations, for instance, a pronouncement of crowd congestion, anomaly detection, and dominant pattern analysis, etc., which may assist in avoiding unnecessary crowding or clogging, and tragic incidents. In high-density crowded places, congested situations can cause crowd tragedies emanating from the maximum density and uneven flow of crowd (anomalous). Our purpose of this study is to model high-density crowd scenes. Here, the term model includes tasks of segmenting crowd motion patterns or flows and detection of abnormalities present in the scene using computer vision and machine learning techniques.

1.2.1 Crowd Flow Segmentation

The flow we are dealing with can be defined as a prevailing path generated by moving crowds in a video. A video can have multiple flows without having the details of the number of flows and the location of each flow. This incomplete information makes the problem of flow segmentation more challenging. The difficulty of a computer vision system is capturing precise motion information, which is depending upon motion representation. The required motion representation should

produce long and consistent trajectories. These trajectories can then be used to define the flow for the entire video.

Usually, optical flow computes pixel-wise flow between frames [Lu et al. (2010); Brox and Malik (2010)]. There are some intrinsic complications in the direct handling of optical flow for motion representation. One is, it generates uncertain outcomes on the borderlines of overlapping flows and acts poorly when the object's movement is prolonged. Other is, long-range spatio-temporal motion samples required in numerous applications are not covered by optical flow. Normally, It is difficult to attain full trajectories of moving crowd in high density. Hence, a concept of tracklet is used to obtain complete trajectories and to capture short-term motion. A tracklet, obtained by the tracker for the object is a section of trajectory. In general, it is challenging or impractical to detect, segment, and track objects in high-density crowd scenes.

To model such high-density scenes, we adopt a holistic approach. The holistic approach considers the crowd as a single entity rather treated as a collection of individuals as in object-based approaches.

The holistic consideration handles the issue of occlusion present in a highly dense crowded scene. These techniques aim to acquire global level details and discard the local ones like main crowd flow is processed while an individual going opposite to flow is discarded.

In this thesis, we present a crowd flow segmentation approach by following an unsupervised paradigm. The approach develops an active contour-based trajectory clustering algorithm. Since the crowd is not present in the whole frame, some portion of the frame is occupied by buildings, walls, trees, etc. Therefore, the active contour approach segments the foreground crowd region from the entire frame to minimize further tracking. Tracking of individuals in the high-density crowd is impractical. We consider a block-level structure of active contour region for trajectory extraction and exploit the essential temporal information depicted in trajectories to handle the complex flow patterns in a better way. We present a new methodology for trajectory representation and feature selection that considers the shape, location, flow direction, and density of trajectories. The extracted trajectories are partitioned into K-primitive clusters using the standard K-mean algorithm and further merged by our presented clustering algorithm. This way trajectories are clustered which have the same flow patterns. It is observed that different trajectory clusters comprise one flow segment. This kind of situation occurs when people with the same flow direction comes into the scene at a dif-

ferent time. To address this issue, we introduce a slightly different version of the density-based clustering technique and segment the crowd flow based on spatial overlapping, direction, and location of clusters. At this point, we obtain physically and dynamically meaningful segments which tell us about how much flow is heading in which direction. Further, each segment is analyzed and, if the density of flow increases beyond a certain threshold, an alert is issued for safety concerns.

The second goal of this crowd scene analysis is the detection of the anomalous scene, which is discussed in the following subsection.

1.2.2 Anomalous Scene Detection in Crowd

Anomalous activities are usually categorized as an outlier detection problem. Existing literature on anomaly detection can be classified as global and local anomaly detection. The global detection identifies whether the scene has an anomaly or not, and the local one tells about the place of an anomaly. Our approach falls into the global anomaly detection method with the aim to detect the anomalous scene, and precisely find out the origin and finishing point of anomaly occurrence in the scene.

Anomalous scene detection in dense crowd consists of two crucial problems. One is severe occlusion between individuals which can be handled by considering the crowd at the holistic level as a single entity. Another issue is the behavior description. A crowd scene contains thousands of behaviors, and it is nearly impossible to describe each of them. Hence, designing a general technique for crowd behavior detection is a cumbersome task. The description of the behavior in this framework can be described as the direction and density of the crowd, whereas anomalous behaviors are the transition which is usually not expected to take place. Here, crowd direction depicts crowd behavior like where the crowd is heading for. Furthermore, through the density information, crowd congestion or bottleneck problem can be identified, if the density of certain place increases beyond a threshold, it is also a sign of an abnormality in the scene.

In this dissertation, we developed an approach for anomalous scene detection that performs statistical analysis on the oriented tracklets of moving crowd. We adopted the two statistical parameters as entropy and temporal occupancy for anomaly detection. The entropy reflects that infrequent events are more informative than frequent ones. Entropy provides a degree of uncertainty or randomness. Higher disorder or chaos leads to higher entropy. And, the temporal occupancy measures

the area occupied by the crowd over time. We can judge the abnormality of a scene concerning massive deviation in both measures. If entropy and temporal occupancy increase beyond a certain threshold, that means something anomalous is happening. Therefore, an alert is issued to prevent potentially dangerous situations. Our experiments are conducted on three datasets: UMN [UMN], UCF Web [Mehran et al. (2009)], and Violent Flows [Hassner et al. (2012)]. The UCF Web and Violent Flows datasets are complementary to the presented approach as they consist of dense crowd scenes of real-world applications such as the marathon, stadium, stampede, and political rallies, etc.

1.3 Motivation

Due to mass gatherings at public places, chances of crowd-related disasters has increased like a stampede, clogging, over-crowdedness, etc. In the recent past, many crowd calamities had occurred due to lack of crowd management, monitoring strategies, and the way our public spaces are constructed. Crowd analysis has become an essential task to provide safe and secure environments in public places. In crowd scene analysis, the focal point is object detection, counting, tracking, and behavior recognition. The conventional techniques are not recommended for high-density crowd scenarios having severe occlusions, small target size, low resolution, and extremely cluttered background. In such scenes, undetected abnormal deeds might lead to undesirable conditions. A crowd has both dynamics and psychological characteristics, and it is very difficult to analyze the behavior and model the dynamics of a crowd at a proper level. Crowd analysis has a large number of real-world applications such as

1. Crowd Management: It helps to develop management schemes to prevent adversities arising due to the crowd for ensuring public safety.

2. Public Space Design: It can give former guidelines in public space designing by guarantying comfort levels and safety measures in the structuring of railway stations, shopping malls, and airport terminals, etc.

3. Visual Surveillance: It provides an automatic procedure to detect the anomaly and generates an alarm for emergency evacuation.

4. Intelligent Environment: This environment can be used to make a wise judgment on how to divert crowd based on their activities at overcrowded places.

5. Entertainment: With profound knowledge of crowd phenomena, the development of mathematical models can give a more precise simulation, which can be exercised in film industries and computer games. It can also be used to produce crowd videos with realistic behavior.

There are places of public domain where crowd analysis has broad applications to detect potential risk due to the large gathering and to control the gathering from being getting overcrowded such as at sports stadiums, protests, demonstrations, marathons, rallies, political speeches, shopping malls, religious places, railway stations and music concerts, etc. which are described by gatherings of hundreds to thousands of people. The public space design and intelligent surveillance is the main scope of this research work. The goal of this thesis is to develop a robust method for crowd density estimation using computer vision approach and to design an efficient algorithm for crowd scene understanding which leads to segment dominant flow pattern by trajectory clustering and detect anomalous scenes by oriented tracklets and generate a quicker response to the emergent situations.

1.4 Research Gaps and Objectives

The computer vision research presented in this dissertation targets two applications of crowd analysis which are public space design and intelligent surveillance. The outcomes of this research can be used to provide guidelines for public space design by estimating the reckoning of customers or commuters that frequently travel through these places. This work also analyzes and segments the crowd flow which assists in avoiding unnecessary crowding or clogging and tragic incidents. Moreover, we detect potentially dangerous situations (anomalous scene) to prevent accidents along with to aid evacuation plans essential in the unlikely events.

Most of the existing techniques are designed for handling a crowd of less than hundreds of individuals, above this the existing literature fails. For crowd density estimation, detection-based methods like head, face, and full-body detectors fail to detect a person in the excessively dense crowded scenes. Any detection or single feature method is not appropriate enough to give a precise crowd density due to severe occlusion, low resolution, perspective, and foreshortening. It is required to formulate a robust method for density estimation in a high-density crowd where each individual occupies a few pixels per target with severe occlusion. The crowd scene analysis depends on motion cue merely between consecutive frames that do

not capture substantial temporal motion. The tracking of an individual is impractical, so we need to consider the whole crowd as a single entity. To segment crowd flow, existing methods track the interest points in the entire frame, though the crowd is not distributed in the entire frame region (i.e., the background contains trees, sky, building walls, etc.). Therefore, only the foreground (crowd region) needs to be tracked to minimize the tracking. Also, in anomalous scene detection, existing methods have a high probability of false alarm generation because they consider only single frame for change detection.

Given the above shortcomings, this research work focuses mainly on the following objectives:

1. Study recent issues and challenges in crowd density estimation, crowd flow segmentation, and anomalous scene detection so that they can be optimized and applied to real-life applications.

2. It has been observed that no single feature or detection method alone is appropriate for density estimation in the dense crowd. Therefore, a multi-feature based method is needed to provide an accurate density count in the high-density crowd.

3. We aim to take into account the challenges offered in dense crowd images like occlusion, foreshortening, and local geometric distortions.

4. Automated analysis of dense crowds requires crowd flow analysis. To find the dominant crowd flow pattern, robust tracking is required to maintain the track of crowd flow. With the recent advancements in computer vision, it can be claimed that there are feasible solutions available for addressing the robust tracking of single or multiple targets. But high-density crowd tracking with severe occlusion is very cumbersome tasks in computer vision. Our objective is to develop and optimize feasible solutions for crowd flow segmentation in the presence of various tracking challenges such as occlusion, short or corrupted trajectory.

5. The performance of crowd flow segmentation mainly depends on the accurate selection and representation of the trajectory feature of the moving crowd. We aim at robust selection and representation of trajectory features.

6. To develop an efficient algorithm for crowd flow segmentation while optimizing the tracking.

7. To generate a method for crowd scene behavior classification (normal or abnormal) by minimizing the false alarm generation.

8. Evaluate the presented methods on benchmark datasets and compare the results with state-of-art approaches.

1.5 Contributions

Our contributions to this study are to create a crowd analysis system which considers the tasks of crowd density estimation, crowd flow segmentation and anomalous scene detection in high-density crowd.
The study summarizes as:

1. Explore the literature work to find various steps involved in crowd analysis.

2. This work presents a texture based multi-source approach for crowd density estimation which includes Fourier analysis, wavelet, HOG and SIFT techniques.

   (a) All these techniques are sequentially applied in local neighborhoods (patch-level) at multiple scales to avoid the issues of foreshortening and local geometric distortion arising from irregularity in the observed textures rising from dense crowd images.

   (b) Simultaneously, the presented method also maintains the smoothness and consistency among neighboring patches by employing Markov Random Field.

3. We present an unsupervised approach for crowd flow segmentation based on trajectory clustering in active contouring specific to high-density scenarios.

   (a) Active contouring segments foreground (crowd region) with the aim to minimize tracking that takes segmented foreground regions as input to select features for further tracking.

   (b) We also continuously take into account the newly appearing moving objects in a video by revising our existing set of tracker constituents after processing a temporal window of frames.

4. We present a new methodology for trajectory representation taking into account the shape, location, flow direction, and density of trajectories.

5. An automatic approach is presented for anomalous scene detection in crowd videos by considering the statistical analysis of oriented tracklets in terms of entropy and temporal occupancy.

6. Validation of our approach of crowd density estimation and crowd scene analysis on benchmarks datasets with remarkable performance compared to existing methods.

The outline of the thesis is visualized in Figure 1.3. Chapter 2 presents a review of existing literature on crowd analysis in terms of crowd density estimation, flow segmentation, and anomalous scene detection. Chapter 3 explains the presented crowd detection approach using the Skin Color Model and Histogram of Oriented Gradient and details its failure cases in a dense crowd. This chapter also provides a detailed description of the presented crowd density estimation approach using texture-based multisource approaches. Chapter 4 discusses a contouring based trajectory clustering approach for crowd flow segmentation and, Chapter 5 describes an algorithm for anomalous scene detection in the high-density crowd. Chapter 6 highlights the concluding remarks upon this dissertation with a discussion of the identified directions for future work.

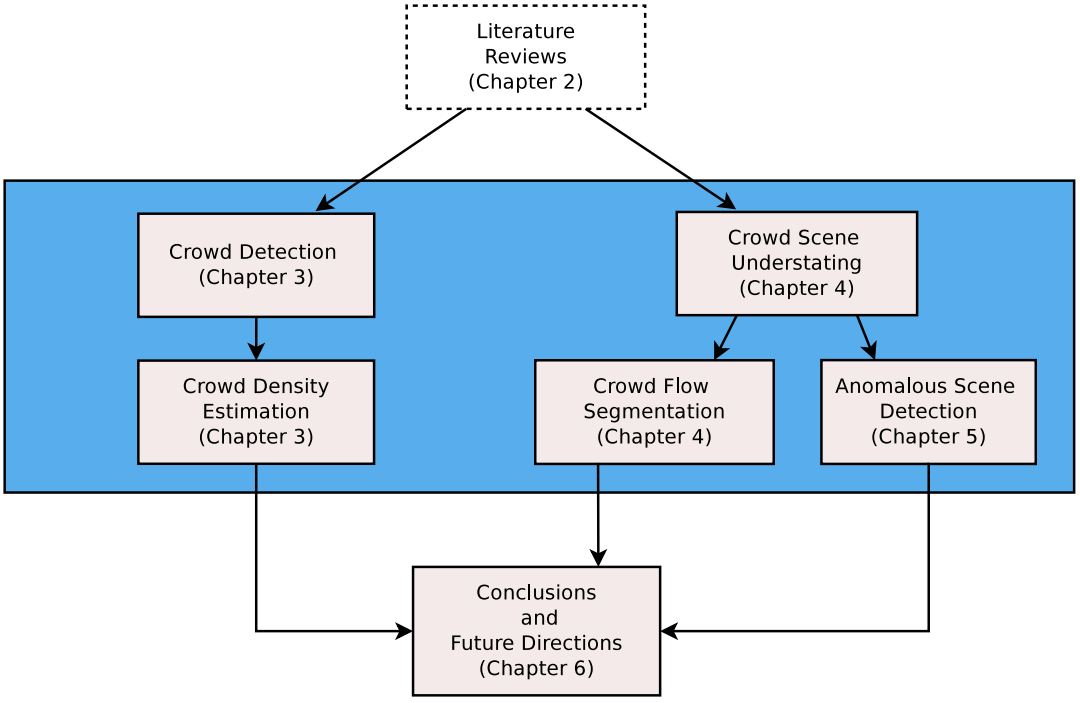

Figure 1.3: Visualization of the thesis outline

Chapter 2

LITERATURE REVIEW

Nowadays, crowd analysis has gained increasing attention along with the increasing concern on public security and safety. Crowd analysis may cover a wide range of research topics, and a lot of literature is available for crowd detection, density estimation, tracking, and anomaly analysis applied to crowd analysis. This thesis mainly focuses on crowd density estimation and crowd scene behavior analysis in terms of crowd flow segmentation and anomalous scene detection. In this chapter, we report a survey of those researchers that share a mutual interest in our work. However, for presenting a complete review, we also include some similar approaches and techniques used for the same task, though we have not used all of them in this work. We first discuss a broad spectrum of crowd analysis approaches proposed in the literature. The structure explained in the previous chapter includes three main tasks for crowd analysis, which includes crowd density estimation, flow segmentation, and anomalous scene detection. This chapter covers recent advances and research related to each of these tasks.

2.1 Crowd Density Estimation

Crowd density estimation is one of the essential tasks in crowd analysis, it counts the number of people in the given image or video. Generally, the techniques of crowd density estimation are broadly categorized into two main approaches: detection based approach and feature-based approach. This section presents an overview of each of the approaches, with specific attention on the feature-based approach that has shown to be effective in dense crowd scenarios.

2.1.1 Detection-based Approaches

Detection based approaches detect occurrences of the pedestrian by examining the image using a detector, which learned with local image features. Most often, detection is executed either in the monolithic based or parts-based style.

Monolithic Detection: Monolithic detection is an instinctive technique to count the number of people in a scene via detection. Monolithic detection approaches [Dalal and Triggs (2005); Leibe et al. (2005); Li et al. (2008)] incorporate conventional pedestrian detection techniques which train a classifier using several features like haar wavelets [Viola and Jones (2004)], histogram oriented gradients [Dalal and Triggs (2005)], edgelet [Wu and Nevatia (2005)] and shapelet [Sabzmeydani and Mori (2007)] etc. as shown in Figure 2.1(a).

The choice of classifier introduces a powerful effect on the speed and accuracy of detection. Non-linear classifiers viz. RBF SVM provides adequate accuracy of detection but at a slow pace. Therefore, linear classifiers like boosting [Viola et al. (2005)], linear SVMs, or Random Forests [Gall et al. (2011)] are widely used. Afterward, pedestrian candidates are detected by applying a trained classifier over an entire image in a sliding window manner. The monolithic detector of the entire body can generate adequate detections in a low-density crowd but are contrarily affected by the presence of occlusion and scene clutter in high-density crowds.

Part based Detection Partial occlusion problems can be resolved to some extent by using part-based methods [Wu and Nevatia (2007); Li et al. (2008); Felzenszwalb et al. (2010)]. In this, a classifier is designed for a specific body part (like head, shoulder) for estimating the people count in a designated region, as shown in Figure 2.1(b).

Comparing with monolithic detection, part-based detection releases the strict assumption about the visibility of the complete body in the crowd scene, thus it is comparatively more robust in crowded scenes. The parts-based detectors reduce the occlusion-related problems but not suitable for highly dense crowds and cluttered backgrounds. However, these kinds of methods are not convenient for the kind of datasets we deal with, because, in highly dense crowd images, an individual occupies very few pixels. Therefore, no human, face or even head are visible, which further gets difficult because of severe multiple occlusions and cluttered background.

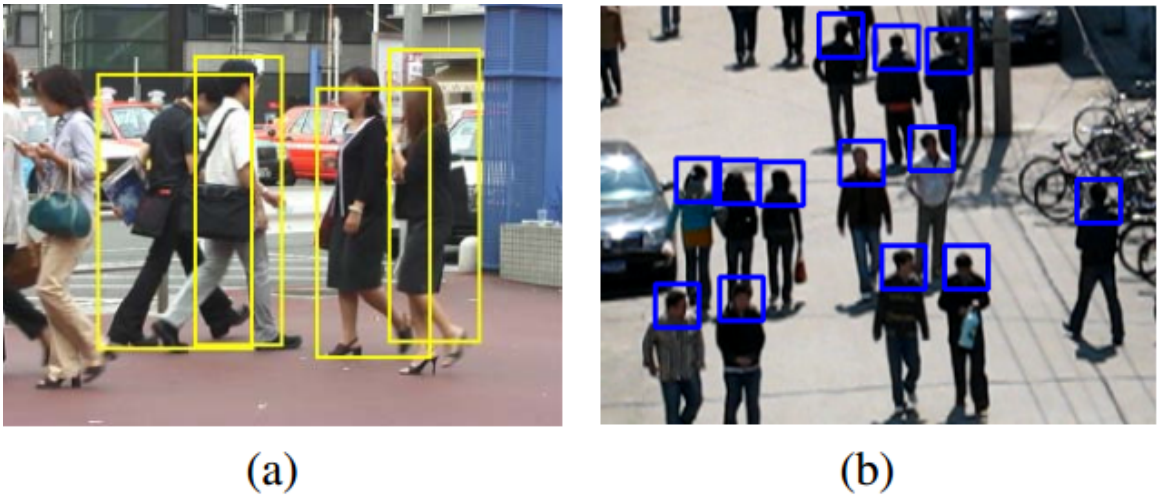

Figure 2.1: Results of pedestrian detection by (a) monolithic detection [Leibe et al. (2005)], (b) part-based detection [Li et al. (2008)].

2.1.2 Feature-based Approaches

To overcome the issues of detection based approaches, researchers explored the feature-based approaches which establish a relationship between extracted features to their counts in a given image. These approaches are divided into two parts: (i) low-level feature extraction, and (ii) regression modeling. Regression methods consciously abstain from individual segmentation and feature tracking. These methods calculate the crowd density on the bases of the holistic illustration of crowd patterns. These approaches neither incorporate segmentation nor tracking. Due to this, regression methods become more practical for a high-density crowd where detection and tracking are extremely limited. Thus, feature-based methods are more efficient than detection based methods as detection of features is easier than the detection of individuals.

Several features of foreground pixels have been widely adopted like foreground area [Davies et al. (1995); Marana et al. (1999); Ryan et al. (2009); Hou and Pang (2011)], texture features [Rahmalan et al. (2006); Chan et al. (2008)], histogram of oriented gradient [Chan et al. (2008); Ryan et al. (2009)] to calculate the density using regression functions such as linear [Davies et al. (1995)], neural networks [Cho et al. (1999)] and Gaussian process [Chan et al. (2008)]. Most of these methods have given a linear mapping between the foreground features and the number of people in the scene. But, this mapping performs poorly in high-density crowd environment due to the complexity in crowd scenes.

Various methods have been developed in the recent past to mitigate the conse-

quences of the perspective problem. For instance, [Paragios and Ramesh (2001)] proposed a geometric factor to add weight at pixels as per its location on the ground plane.

[Ma et al. (2004)] introduced a geometric correction to take all different distance objects at the same scale and perspective mapping to weight all extracted features of an image.

Moreover, occlusion problem can be mitigated by using various features such as histogram of oriented gradient [Kong et al. (2005)], edge features [Davies et al. (1995)], etc. Also, these methods suffer in complex crowd scenes due to several reasons. For instance, the Edge-based features can give inaccurate results due to complex and cluttered backgrounds and also very time consuming to extract such big features, foreground separation becomes more complicated in a dense crowd.

Therefore, few techniques have been designed which make use of local features to tackle these issues and minimize the need for training data. An image broadly offers two varieties of features; pixel-based features and texture-based features, which are used for encoding low-level information. These methods are detailed in the upcoming sections.

2.1.2.1 Pixel-based Analysis

Pixel-based methods rely on locally computed features for estimation of crowd density. These methods utilize low-level features and maintain a linear mapping with foreground pixels and human count. [Velastin et al. (1994)] and [Davies et al. (1995)] have introduced most popular crowd density estimation methods in computer vision. They introduced two automatic approaches based on pixel-level information. The former method extracts pedestrian pixels by examining a three-pixel neighborhood of the difference image. The later applies a fast three-pixel-neighborhood edge detector to the image to obtain edge magnitude and more enhancement is done by thinning the edges as shown in Figure 2.2 with the help of Kalman filtering approach. They proposed a linear model that maintains a correspondence between the resultant binary image and the number of people.

[Cho et al. (1999)] introduced a feed-forward neural network (FFNN) based density estimation system. In their approach, they performed a global learning algorithm by combining least squares methods with random search, simulated annealing,

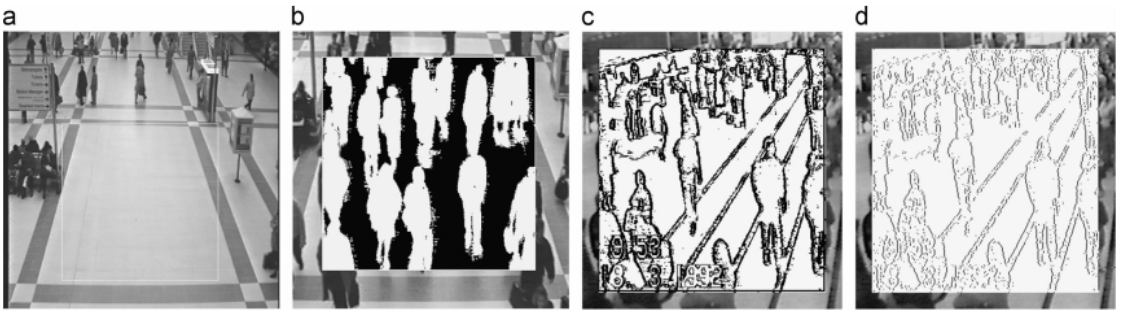

Figure 2.2: The results of [Velastin et al. (1994)] using static background image (a) Reference image, (b) background removal, (c) edge image, and (d) thinned image

and genetic algorithm. Their work concludes that hybrid of least-square and the random search algorithm performs the best among all other hybrid combinations.

[Ma et al. (2004)] also proposed a crowd density estimation method that derives a mathematical relation of geometric correlation and maintains a linear relationship between foreground pixels and people to count human objects. This method is only suitable in the absence of occlusions among the people.

[Choudri et al. (2009)] proposed a pixel-based robust selective background model for crowd counting. Their method detects only pedestrian pixels rather than considering the whole foreground pixels. This method reduces miss detection with the help of more potent people counting classifiers even after being slow or stationary. [Hussain et al. (2011)] introduced an automatic density estimation system specifically for Masjid al-Haram crowd scene. They first employed background subtraction and then carried out a relationship with foreground pixels and density count. They performed supervised training to categorize the crowd density into five different levels ranging from very low to very high. This system can detect the crowd with almost 100% accuracy but in very low to low crowd density. The performance of the system decreases in a high-density crowd, mainly due to severe occlusion. Therefore, the pixel-based methods can not be applied in case of severe occlusion in a dense crowd. Table 2.1 summarizes pixel-based techniques that are commonly used for crowd density estimation.

2.1.2.2 Texture-based Analysis

The texture is one of the most prominent image characteristics that has gained attention from many researchers. The texture is considered as an efficient feature to deal with an extremely dense crowd. It primarily focuses on crowd density

Table 2.1: Pixel-based crowd density estimation approaches and their learning models.

Author	Year	Image Feature	Learning Model	Site
Davies et al. (1995)	1995	Foreground pixels and edge detection	Linear	Indoor
Cho et al. (1999)	1999	Foreground pixels and edge detection	FFNN	Indoor
Ma et al. (2004)	2004	Foreground pixels	Linear	Indoor
Choudri et al. (2009)	2009	Foreground pixels and edge detection	Linear	Outdoor
Hussain et al. (2011)	2011	Foreground pixels and edge detection	BPNN	Outdoor

Figure 2.3: Crowd density classification ranging in very low, low, moderate, high and very high density levels respectively [Marana et al. (1998)].

estimation instead of individual counting in the crowd scene. Various texture-based approaches have also been proposed for density estimation in a dense crowd. [Marana et al. (1998)] and [Marana et al. (1999)] observed that dense crowd images depict fine texture whereas sparse crowd images show coarse texture.

For density estimation, most of the texture-based methods incorporate GLDM, Fourier analysis, Wavelet analysis, SIFT, LBP and fractal dimension, etc. In their work, Haralick features known as Gray Level Dependence Matrices (GLDM) are used to extract crowd density features. Then, these features are supplied to the Self-Organizing Mapping (SOM) neural network to categorize the crowd images into five density levels, as depicted in Figure 2.3. They combine the work of [Marana et al. (1999)] with Minkowski Fractal Dimensions (MFD), which takes an advantage over GLDM and Fourier spectrum. This work is not feasible for real-world crowd scenes because it demands a smooth background which should be free from objects.

In further work of [Marana et al. (1998)], the performance of four different methods like GLDM, straight-line segments, Fourier analysis, and fractal dimension is compared for texture analysis. Furthermore, three different classifiers, namely

SOM neural network, statistical Bayesian classifier, and fitting based approach are compared. They get the best results by combing GLDM with the Bayesian classifier.

Following the similar concepts of [Marana et al. (1998)], [Xiaohua et al. (2006)] introduced a fusion of multiscale texture analysis and SVM. This hybrid method outperforms the work of [Marana et al. (1998)] and [Davies et al. (1995)] with respect to the computational complexity, but it decreases the attainment when the crowd becomes non-uniform. [Brostow and Cipolla (2006)] and [Rabaud and Belongie (2006)] count moving objects by calculating coherent motion pattern. [Kratz and Nishino (2009)] and [Wang et al. (2007)] also proposed the coherent motion patterns, but these patterns are not explicitly applicable for crowd density estimation. These techniques are only applicable for moving scenes with a high frame rate. These are not appropriate for still crowd images and videos that have negligible motion viz. speeches, concerts, etc.

Another technique that relies on regression-based feature [Cho et al. (1999)] provides a direct relationship between local features and density estimation. This regression parameter is learned for the entire image. They have the implicit postulation that the density is uniformly distributed in the entire image. This hypothesis is fallacious in most of the realistic crowd scenes due to camera viewpoints and high variable density.

[Lempitsky and Zisserman (2010); Ma et al. (2010)] proposed a cell-based method for regression in which images are partitioned into small cells. Their aim of image partitioning is to remunerate the localized geometric distortions by cause of perspective. The primary issue with this approach, however, is to ignore spatial consistency constraints as information among local neighbors is unshared. Recently, [Chen et al. (2012)] claim that information shared among local neighbors provides more precise and robust results of crowd counting. They offered a single multi-output framework, but it was limited to the scenarios with a few tens of people.

Furthermore, the great success of Convolution Neural Network (CNN) in innumerable computer vision applications has encouraged researchers to take advantage of their strengths for learning nonlinear functions from crowd images to their corresponding density maps or counts. Numerous CNN-based methods have been introduced in the recent past. [Wang et al. (2015)] and [Fu et al. (2015)] were among the first ones to adopt CNN's for estimating the density of the crowd. [Wang et al. (2015)] have given a CNN based deep learned regression model for

dense crowd counting. They adopted AlexNet network architecture [Krizhevsky et al. (2012)] in which a single neuron layer takes the place of a fully connected layer of 4096 neurons to estimate the crowd density.

[Zhang et al. (2015)] proposed a cross scene counting method and explored that the performance of existing CNN based methods decreases remarkably when evaluated on an unseen video (different from the training dataset). Another similar work done by [Hu et al. (2016)] adopted a CNN for extracting features and count individuals in the crowded scenes with mid-level or high-level densities. A comprehensive survey of crowd behavior analysis using convolutional neural networks can be found in the research article written by [Tripathi et al. (2019)]. We conclude that for cross scene analysis, a massive amount of datasets are required, but the availability of such a large number of crowd datasets is minimal. Moreover, augmentation of the training dataset is performed by replicating the images which lose the variability of scenes. To summarize, deep learning-based techniques could play better but with the requirement of expensive GPUs and the massive amount of datasets. Furthermore, to help the readers in understanding the present day challenges of crowd density estimation, we have depicted the tabular comparison among different techniques as shown in Table 2.2.

Author	**Approach**	**Regression or Learning Model**	**Limitations**	**Dataset**
Davies et al. (1995)	Foreground pixels, edge detection	Linear	Relied on static background model, effects of perspective not apparent	–
Marana et al. (1999)	MFD	Linear regression	Classifier not able to classify due to partial fractality, limited discriminate power, time consuming, sensitive to noise	Liverpool Street Railway Station, London
Lin et al. (2001)	HWT	SVM	Works only when head contour visible, occlusion	Random images
Siebel1 et al.(2003)	Motion detection	SVM	Illumination effects, installation and maintenance cost	Manually captured
Leibe et al.(2005)	Chamfer Matching	SVM	Train only 44 sequences of 35 different person with two background	No dataset
Rahmalan et al. (2006)	MFD, GLCM, TIOCM	SOM	Highly sensitive to external change, illumination, low density	Manually recorded
Kong et al. (2006)	MoG adaptive background modeling, edge orientation map	Linear fitting, neural networks	Occlusion, complex background	No dataset
Chan et al. (2008)	Segment feature, internal edge feature, GLDM	Feature based regression	Outlier due to lack of training data	UCSD
Ma et al. (2008)	LBP texture analysis	K-mean clustering	Perspective distortion, occlusion, clutter background	4 scene

Ryan et al. (2009)	Blob segmentation	NN	Only group segment, poor segmentation and low scalability	UCSD
Dong et al. (2010)	Texture feature-GLDM, HOG, GGDM	Clustering	Only indoors, more advisable to use contours or shapes for crowd estimation	Subway surveillance videos
Ma et al. (2010)	Density classification-GOCM	Cluster classification	Low density, occlusion, only edge detection and perspective distortion	Manually collected-hall, square, subway, platform
Santoro et al.(2010)	Optical flow	Density based clustering	Only 2D coordinates for motion points, distance between motion points is not accurate, low density	PETS2009
Hsu et al. (2011)	DCT-frequency analysis	SVM	Limited performance, occlusion, very few people	No-dataset
Srivastava et al. (2011)	Crowd Flow: GLCM, foreground pixel,	Weighted pixel summation	Sensitive to number of density level, occlusion,	UCF, UCSD
Yang et al. (2011)	SST-LBP	SVM	Does not work in restricted flow or dense flow	PETS 2009
Mousavi et al. (2012)	Uniform LBP	Unique model histogram	Fails in high density, occulation	Soccer field video sequence
Chen et al. (2013)	Foreground segmentation, edge and local texture features	SVR	Low density, poor segmentation, illumination	USCD and Mall dataset
Fradi et al. (2013)	Density map: FAST, RLOF	Gaussian symmetric kernel	Low density count	PETS 2009 and UCF
Herrmann et al. (2013)	Local Fourier analysis	Regression method	High false positive	Aerial-manually collected images
Idrees et al. (2013)	SIFT, Fourier, DPM	SVR, MRF	Implicit assumption of uniform density at patch level	UCFCC
Meynberg et al. (2013)	FAST corner detector	SVM	Local inconsistency	Aerial images
Tian et al. (2013)	LCSSF (local color self Similarity Feature)	SVM	Fails in 5-6 people, occlusion	MIT, CAISIA GAIT pedestrian
Forsyth et al. (2014)	Object detection-DPM	SVM	Labeling is required, does not work in crowd	PASCAL
Handte et al. (2014)	WLAN-devices	K-NN	Should reduce the operational costs of the network	Public transport system-Madrid
Khan et al. (2014)	Forground segmentation, motion flow segmentation	K-means	Low precision	UCF

Maddalena et al. (2014)	Self-organizing background subtraction	k-NN	Low accuracy at large crowd	EPFL, PETS-2009
Perko et al. (2014)	HOG	SVM	Scale variant	Manually collected
Zhang et al. (2014)	Forground pixel extraction	Statistical regression	Not suitable for crowd, occlusion	No dataset
Fradi et al. (2015)	Robust Local Optical Flow	Gaussian kernel	Only tested on low density, perspective effects	PETS, UCF, UMN
Idrees et al. (2015)	Combination of Parts-DPM	Latent SVM	Low resolution, lower limit of DPM 23×23 pixels	UCF-HDDC
LE et al. (2015)	Gabor filter	SVM	Prior knowledge of people behavior is required, Group	PETS
Wang et al. (2015)	DNN	Deep-CNN	Can feed deep model with negative samples, large number of training sets required	UCFCC
Zhang et al. (2015)	Foreground segmentation, GLCM	Learning distribution regression	Required labeling on training data	UCSD and Mall
Meynberg et al. (2016)	BoW, Gabor filter bank	SVM	Low resolution, illumination	Aerial images
Hu et al. (2016)	NN	ConvNet	Need to include temporal and multiresolution information	AHU-CROWD
Siva et al. (2016)	HoMG, background subtraction	SVM regression	Low density crowd	PTES, QUT, Mall

Table 2.2: Comparative analysis among the literature of crowd density estimation

2.2 Crowd Scene Behavior Analysis

Nowadays, crowd scene behavior analysis is an active research topic in computer vision and pattern recognition. In the past few years, various state-of-the-art techniques are introduced to tackle crowd analysis and behavior understanding. Generally, the temporal information like direction, velocities, and unusual motions is used to analyze the behavior of the crowd. Approaches for crowd behavior analysis can be broadly classified into object-based and holistic approaches. The object-based techniques deal with crowd considering it as a collection of individuals while in holistic approaches, a crowd is considered as a global entity to judge the behaviors of the whole scene.

2.2.1 Object-based Approaches

Object-based crowd behavior analysis is inferred by performing traditional detection and segmentation approaches on the individuals. For instance, the detection or tracking of a single person moving against the dominant flow could be an anomalous activity. Object-based methods can solve this kind of problem only in low to moderately crowded scenes. These methods face complexity in recognizing individuals in high-density crowd scenes due to a large number of targets, small resolutions, small target size, and severe occlusions, etc. Such factors increase the chances of information loss related to target objects in crowded scenes. Therefore, individual segmentation and tracking are not feasible in the high-density crowd.

To overcome this problem, some of the researchers [Wang et al. (2009); Zhou et al. (2012b)] have adopted low-level features and probability models to analyze the dense crowd instead of focusing on tracking individuals. [Wang et al. (2009)] explored the hierarchical Bayesian model to fulfill many challenging video surveillance jobs like segmenting motions into different activities by utilizing visual features and atomic activities. Their technique neither requires labeled training nor tracking.

[Zhou et al. (2012b)] classify different pedestrian behavior based on the mixture model of dynamic pedestrian agents, which learns the collective behavior patterns on the pedestrian. Once the model is trained by various samples, the mixture model can predict past and future behavior of pedestrians and classify the pedestrian behavior in the crowd scene. But, the mixture model has some shortcomings, for instance, it considers the only affine transforms and faces difficulty in representing complex shapes.

2.2.2 Holistic Approaches

In high-density crowded scenes, individual objects appear unresolved. Therefore, rather tracking individual objects, a top-down approach is followed, which considers the crowd as a single element. This approach handles the issues arising from the occluded crowd, which is not tackled by object-based approaches. Holistic approaches generally extract the coarser-level information globally like main crowd flows. They discard local information, such as an individual moving out of the flow. In highly crowded environments, very few features are extracted from

individuals, and some of them remain undetermined. In such situations, analyzing crowd behaviors without recognizing individual activities is usually beneficial.

In the holistic approaches, a crowd is considered as a global entity to judge the behaviors on the whole scene. These methods inspect the dynamics of the entire crowd scene instead of focussing on individual activities. Therefore, these techniques escape the exercise of individual detection and tracking people, and explore crowd features to analyze the whole behaviors.

This class usually incorporates optical flow field-based approaches [Ali and Shah (2007); Hu et al. (2008); Loy et al. (2009); Luvison et al. (2011); Benabbas et al. (2011) and [Krausz and Bauckhage (2012); Rao et al. (2016)]. [Krausz and Bauckhage (2012)] represented global motion pattern by utilizing an optical flow histogram, which used to detect stampede situations such as Love Parade stampede. [Benabbas et al. (2011)] developed a crowd model based on direction and magnitude and proposed a region-based segmentation algorithm which detects crowd event by learning different motion patterns. [Rao et al. (2016)] developed an optical flow-based probabilistic framework using Riemannian manifolds to detect crowd activities.

In contrast to existing work that depends on motion cue merely between consecutive frames, more advanced approaches are developed that enhanced this cue to larger temporal frames by tracking some salient [Shao et al. (2014); Mousavi et al. (2015)] points or particles advections [Mehran et al. (2009); Mahadevan et al. (2010); Mehran et al. (2010)]. This method results in trajectories that capture more substantial temporal motion, which helps in to analyze crowd motion patterns. [Mousavi et al. (2015)] represents crowd motions by a histogram that encodes magnitude and direction of motion. This is mostly used to recognize an abnormality in the crowded scenes. Various methods have been introduced in the literature on crowd behavior analysis by utilizing motion direction and magnitude of trajectories. These methods only focus on individuals properties.

To consider the collective information about the crowd, [Mehran et al. (2009)] built a social force model (SFM) that detects abnormal crowd behaviors by measuring the interaction force among individuals in the crowd. This method considers the collective information about the crowd and estimates the particle interactions to identify abnormal scenes involving fighting or escaping events. Some of the authors [Mahadevan et al. (2010); Mehran et al. (2010); Yuan et al. (2015)] have also offered similar techniques.

[Mahadevan et al. (2010)] proposed a clustered particle trajectories to detect anomalous events in the crowd. The chaotic invariants of all clustered trajectories are quantified as maximal Lyapunov exponent and correlation dimension.

[Mehran et al. (2010)] proposed an interaction energy potential function to model the group activities based on social behavior. [Yuan et al. (2015)] proposed a structural context descriptor (SCD) that considers how each tested target is related to other individuals present into the scene. The SCD descriptor detects and localizes the abnormality by online spatio-temporal analysis.

In recent past, few researchers focus on particle advection approach [Mehran et al. (2009); Gu et al. (2014)]. In this approach, a grid of particles is randomly distributed on the frame to represent crowd individuals and which are advected along with the optical flow. [Gu et al. (2014)] proposed a particle entropy approach for analyzing the crowded scenes.

In their particle advection scheme, particle resembles pixels, and the pixels are hard to differ with their neighborhood, which results in corrupted tracklets. This issue is addressed in this dissertation by considering the spatio-temporal interest points to be tracked. We also filtered out the short length trajectories to not affect the persistent crowd flow direction. In the following sections, we discuss the literature of holistic based crowd flow segmentation and anomalous scene detection methods.

2.2.3 Crowd Flow Segmentation

Flow segmentation is one of the most explored research problems in crowd analysis. Various techniques have been introduces in the past to address the issues of identifying dominant flow pattern in high-density crowd videos [Ali and Shah (2007); Wu et al. (2009a); Mahadevan et al. (2010); Rodriguez et al. (2011b); Wu and San Wong (2012); Kruthiventi and Babu (2015)].

[Ali and Shah (2007)] introduced a crowd flow segmentation approach influenced by the fluid dynamics principle. They proposed a Lagrangian dynamics-based approach, in which a grid of particle is placed over the frame and advected along with the optical flow field to generate trajectories of these particles. In the crowd flow, the Lagrangian coherent structures (LCS) are detected to separate the flow boundaries by using finite-time Lyapunov exponent field. Further, they enhanced

their work to detect instability of a particular region by finding the abnormality in the segmented flow field.

[Cheriyadat and Radke (2008)] reviewed several computer vision algorithm and proposed an optical flow-based algorithm to develop a system for the automatic detection of dominant crowd motion patterns. To handle spatial and temporal changes efficiently, [Mehran et al. (2009)] also utilized the locations of individual particles that passed through a particular point.

[Rodriguez et al. (2011b)] proposed a supervised method to analyze the crowd behavior pattern learned from large datasets. From the unsupervised perspectives, an optical flow-based region growing scheme was developed by [Mahadevan et al. (2010)] to segment crowd flows. They also proposed a flow segmentation approach, which is based on fuzzy c-means clustering [Wu et al. (2009b)].

Another local translation domain segmentation model [Wu and San Wong (2012)] has been developed by computing optical flow only at salient points. [Kuhn et al. (2012)] had developed an automated crowd behavior analysis framework with the combination of optical flow and Lagrangian analysis of time-dependent vector fields. The outcomes of motion analysis are also useful to recognize various crowd behaviors like circle formation, blockage, bottleneck, etc. depending upon the detection of accumulation points. These points can be further utilized to track back the crowd motion.

On the other hand, to detect a salient region in a crowd scene, [Loy et al. (2012)] proposed a salient motion-based approach that identifies global anomalous flows by inspecting spectral singularities in the motion field. Similarly, [Lim et al. (2014)] introduced an approach that detects saliency in the crowd by projecting low-level features into global similarity structure; such structure concedes the exploration of inherent motion dynamics.

In the recent past, researchers have addressed this flow segmentation problem in compressed domain [Biswas et al. (2014); Kruthiventi and Babu (2015)] by using a motion vector information. In [Kruthiventi and Babu (2015)], the motion vectors are modeled as conditional random fields (CRF) and obtain the flow segments which have the minimum global energy of the CRF model. CRF is a statistical model useful in assigning the labels to sequential data that defines the probability of an individual label sequence.

In another work of [Biswas et al. (2014)], motion vectors are clustered by employing an expectation-maximization algorithm. [Fradet et al. (2009)] proposed a pixel-

wise segmentation approach, which is based on trajectory clustering to segment motion patterns in a dense crowd.

Our approach handles the flow segmentation problem by exploring the pixel domain where the entire video region (either crowd present or not) is considered to track and further processed to achieve the crowd flow segmentation. Since tracking the entire region is computationally expensive, and also not suitable for real-world applications. We present a novel approach to accomplish flow segmentation in an active contour region by using trajectory clustering, which can handle both linear and intersecting flows of dense crowd scenes.

2.2.4 Crowd Anomaly Detection

In crowded scene analysis, anomaly detection is the main task which has drawn huge attention in recent past [Ali and Shah (2007); Mehran et al. (2009); Kratz and Nishino (2009); Mehran et al. (2010); Cong et al. (2013)]. In spite of enormous efforts, anomaly detection still remains an open research problem both in terms of approaches and problem definition, hypothesis, and targets [Sodemann et al. (2012)].

Crowd anomaly detection methods can be supervised in various ways, such as from labeled data of both behavior class, i.e., normal and abnormal, or from a collection of unlabeled data, with most part normal as an assumption. Based on the scale of interest, anomaly detection can be classified into two categories: (i) global anomaly detection (anomaly exists in the scene or not), and (ii) local anomaly detection (determine the place where the anomaly is taking place) [Cong et al. (2013)].

In global anomaly detection, the whole crowd is considered as a single element. Its key objective is to determine the dominant and (or) anti-dominant patterns of the single event, without considering any particular behavior. For example, the scenarios like congestion or stampede are a convergence of a crowd's locomotion. Due to this, global analysis, instead of focusing on specific behavior, focuses on overall tendencies of the critical mass. Local anomaly detection concentrates on the crowd more precisely. It focuses on the detection of individual behavior, actions among other crowd entities. This task becomes more challenging when the crowd density is too high, as in such cases, occlusion makes the task cumbersome. The detailed description is provided in the following section.

2.2.4.1 Local Anomaly Detection

Often, it becomes necessary to know the place of anomaly occurrence. To target this goal, many local anomaly detection techniques are designed. Popularly used models ranging from crowd dynamics and vision areas are being used for local anomaly detection. These approaches fall under two categories: (i) vision-based approaches, and (ii) physics-inspired approaches. Vision-based approaches are based on visual features to supervise the model and predict the anomaly. Physics-inspired approaches make use of physical models for the representation of the crowd dynamics. It detects the anomaly using different learning methods.

Vision-Based Approach: Nowadays, a lot of machine learning approached have gained success in vision tasks, which is also applicable to local anomaly detection. These techniques extract the visual features and make cluster sets to depict various possible event patterns.

(a) Dynamic Texture Model: The dynamic texture [Chetverikov and Péteri (2005)] is a generative model used spatio-temporal features to detect anomaly in videos. It depicts video sequences as an observation from a linear dynamical system, and shows spatio-temporal stationary properties [Xu et al. (2011)]. As per the recent literature [Mahadevan et al. (2010)], a dynamic texture is better than optical flow for local event detection in crowded scenes. The approach proposed by [Chan and Vasconcelos (2008)] is a mixture of dynamic texture (MDT) for motion segmentation. In their technique, for representing samples from a set of underlying dynamic textures, a set of video sequences is used. The Figure 2.4 depicts the MDT technique from a video subsequence. [Li et al. (2014)] designed a unique detector that can be used for both temporal and spatial anomalies using MDT. This detector is applied over a video sequence that considers both dynamics and appearance using various MDT models. During the training phase, normal patterns are supervised through the MDT model for every scene subregion. A multi-scale temporal anomaly map is calculated based on the negative log probability for each video subsequence, which lies under the MDT of its respective region. In the testing phase, the subsequence of low probability for respective MDT, are categorized as anomalies.

(b) Hidden Markov Model: The HMM is able to consider inherently dynamic nature of the observed features [Cong et al. (2013)].It is useful both in case of video event detection as well as anomaly detection. The framework designed by [Kratz and Nishino (2009)], model local spatio-temporal motion behaviors in

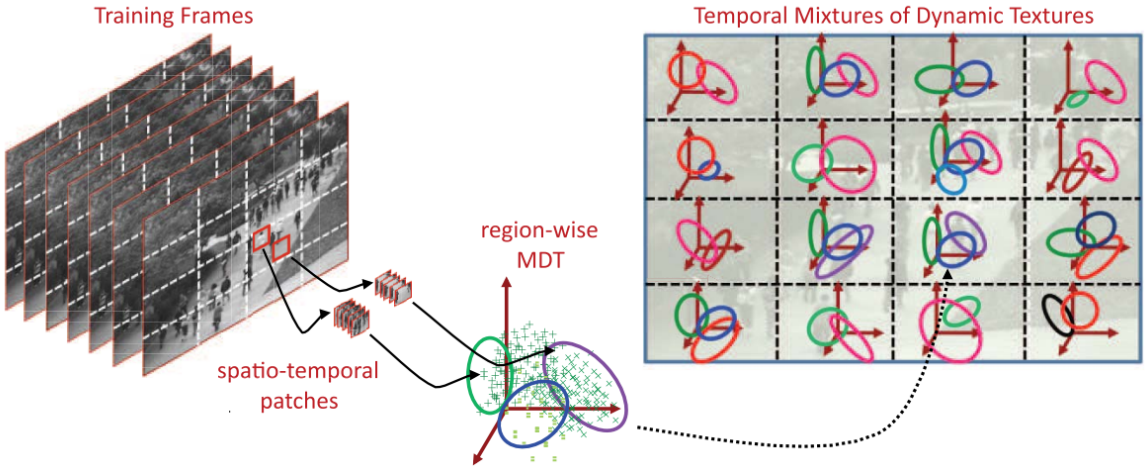

Figure 2.4: This image is reproduced from [Li et al. (2014)]. A mixture of dynamic texture is learned from spatio-temporal patches to detect temporal abnormalities.

a high-density crowd. The training phase of their method maintains a temporal relationship between local motion patterns acquired via distribution-based HMM whereas the spatial relationship is captured by a coupled HMM. During the testing period, abnormal events are identified if it shows some statistical deviations in video sequences of the same scene. It is shown in experimental results that the proposed approach is suitable for analyzing anomalies in extremely dense crowded scenes. Since authors have set up one single HMM for a local area, therefore it will work only for limited normal behavior or specific crowded scenes. Unless the model is retrained, changing the type of normal behavior decreases the detection rate of abnormal behaviors. Similarly, [Wang et al. (2012)] proposed an approach which calculates high-frequency and spatio-temporal (HFST) features using wavelet transformation, for determining the dynamic structure of the local region. Then, the detection of various local anomaly crowd events is performed using multiple HMMs, and individual HMM accounts for a different type of behavior.

(c) Manifold Learning Model: This framework [Thida et al. (2013)] has also been widely used for anomaly detection in a crowded scene. To understand the local motion structure, spatio-temporal Lagrangian eigenmap method is used. In both spatial and temporal domains, a pairwise graph is constructed taking into account the visual context of multiple local patches. Such a process consider local motion patterns for different spatial locations, where similar pattern are placed closely, and different are distant apart. It helps to cluster embedding points and to determine the various motion pattern in crowded scenes. Eventually, abnormal regions are located using a local probability model. The clusters having small data

points or outliers are marked as abnormal.

Physics-inspired Approach: Various physics-inspired models have also been suggested for crowd representation. These model in combination with various machine learning techniques are helpful for anomaly detection. A few examples of these models are a continuum-based approach and an agent-based approach.

(a) Flow-field Model: In order to find out regular patterns, there is a requirement to have a knowledge of how and where crowds evolve with time. It helps to know the place and the way in which crowd motion pattern changes. Based on flow field, [Ali and Shah (2007)] has proposed techniques for motion pattern segmentation which have been further extended for anomaly detection. They developed a finite time Lyapunov exponent (FTLE) field. Its boundary changes in accordance with the variation in the crowd in terms of the dynamic behavior of the flow.

[S. Wu (2010)] designed a method for crowd flow modeling and anomaly detection for both structured as well as unstructured scenes. It begins with particle advection which is based on optical flow. Further, particle trajectories are clustered in order to get representative trajectories. Based on chaotic invariants, chaotic dynamics of all representative trajectories are extracted and quantified. This process is termed as maximal Lyapunov exponent and correlation dimension in the dynamic system. From the obtained chaotic feature sets, a probability model is supervised. Eventually, a query video of a scene is marked as normal or abnormal based on a maximum likelihood estimation criterion.

(b) Social Force Model: This model is popular and widely used in research fields as simulation and analysis of crowds. Using SFM, [Mehran et al. (2009)] designed a technique for detection and localization of abnormal behaviors in crowd videos. Therefore, the framework proposed by [Ali and Shah (2007)] is used for computing particle flows, and their interaction forces are estimated using SFM.

After that, mapping of interaction force into an image plane is done for obtaining force flow for every pixel in every frame. Using spatio-temporal volumes of force flow, normal behavior is modeled. Finally, the frames are classified as normal or abnormal by using BOW. With the help of interaction forces, the region of anomalies in the abnormal frames are localized. Inspired by SFM, laterally, some more methods [Raghavendra et al. (2011a,b)] have been proposed to detect abnormal crowd behavior.

(c) Crowd Energy Model: The crowd has its own characteristics to measure

the crowd dynamics, for instance, local density, velocity, etc. [Yang et al. (2012)] proposed an efficient technique for crowd anomaly detection using the histogram of oriented pressure (HOP). SFM and local binary pattern (LBP) are used to calculate the pressure. Cross histogram is employed to generate the feature vector instead of parallel merging the magnitude histogram and direction histogram. Subsequently, the support vector machine and median filter are used to identify the anomaly.

[Xiong et al. (2011)] also suggested a scheme to identify two typical abnormal activities: pedestrian assembling and running. The scheme is based on the potential energy and kinetic energy. A name, crowd distribution index (CDI) is termed to signify the dispersion, which can afterward determine the kinetic energy. In the end, the unusual activities are identified through threshold analysis. The energy-based model can well represent the dispersion on diverse directions and find moving information and interacting information between individuals. The model functions because obvious differences exist amid normal states and abnormal states in dynamic crowd features. Generally, some threshold-based methods are engaged here, and the threshold is determined analytically when applied to diverse crowd scenes.

2.2.4.2 Global Anomaly Detection

Generally, a crowd scene exhibits a regular motion pattern due to self-organization effects. But, when abnormal events influencing public safety occur, such as stampede, explosions, fires, transportation disasters, and it turns crowd dynamics into a totally different state. Global anomaly detection intends to differentiate the abnormal states of the crowd from normal ones. Correlated methodologies typically have a tendency to identify the changes or events based on the obvious motion assessed on the whole. A global anomaly detection system should correctly find out the starting and end of the events, as well as the transitions between them along with identifying the presence of an anomaly in the scene. Holistic methods for crowd scene behavior analysis considered in Section 2.2.2, such as [Mehran et al. (2010); Solmaz et al. (2012); Su et al. (2013)] can be employed for global crowd anomaly detection. Some works carried out in the global style exclusively for anomaly detection also exist in the literature which is as follows.

[Chen and Huang (2013)] presented an anomaly detection method in which human crowd is considered as a graph and isolated region is represented as a vertex. To efficiently model the topology differences, local and global features are used. These

features are merged as an indicator to identify if any anomaly of the crowd exists in the scene.

Recently, [Wu et al. (2014)] proposed a Bayesian framework for crowd escape behavior identification in videos to directly model crowd motions as non-escape and escape. Crowd motions are categorized by means of optical flow fields, and the correlated class conditional probability density functions are formed based on the field characteristics. Crowd escape behavior can be identified by a Bayesian formulation. Experimental results depicted that the approach is more precise than state-of-the-art approaches in detecting crowd escape behavior. But, this approach is not suitable for high-density crowded scenes, because the high-density crowd behavior varies from the behavior in low or medium density scenes.

2.3 Chapter Summary

In this chapter, we presented a literature survey of fundamental phases of crowd analysis. Here, we have discussed the various techniques used for crowd density estimation, especially crowd texture feature extraction. This step is a critical step in crowd analysis because the result of later steps is primarily dependent on it.

After crowd feature extraction and crowd density estimation, crowd behavior detection is the inherent part of crowd analysis systems. We have surveyed the various existing techniques of crowd flow segmentation. Our review of the study is that effective feature selection in dense crowd tracking is an essential step that may affect the result of tracking significantly and segment the crowd flow pattern effectively.

Next, we surveyed the holistic approaches for detecting the anomalous crowd scene present in a crowd video. These methods consider the whole scene as a single entity which are using labeled data for training of normal and abnormal crowd patterns. We discussed both local and global anomaly detection methods based on situation demands or interests. In this thesis, we adopted global methods because " what is happening is more important than who is doing it".

In the next chapter, our proposed approach of crowd density estimation is explained. It is pertaining to state-of-the-art detection systems, as well as to the algorithms used for crowd detection, considering many of the relevant algorithms provided in this chapter.

Chapter 3

Crowd Density Estimation

This chapter focuses on the problem of crowd density estimation in images of extremely dense crowds. Our aim for density estimation is to predict the number of people in the given image. The techniques of crowd density estimation are broadly categorized into two main approaches: (i) detection based approach and (ii) regression-based approach. Detection is a fundamental step in density estimation. Though, the detection based approaches perform poorly in dense crowds where the human bodies, faces, and heads are severely occluded and occupy just a handful of pixels. But, detection of face and head adds robustness to crowd density estimation by performing well in relatively low-density environments where crowd faces are visible or facing the camera.

In this chapter, we present detection and regression-based crowd density estimation method by implementing a skin color model, Markov random field, support vector machine, and various texture-based features.

3.1 Detection-based Density Estimation

Previous works addressing the crowd counting problem majorly follow the counting by the detection method. Initially, a detection based style was explored for crowd density estimation, which uses sliding window detector [Dollar et al. (2012)] to locate people in the scene. Most often, detection is executed in two styles (i) monolithic-based style, and (ii) part-based style. This method trains a classifier by utilizing various features like HOG, edgelet, shapelet, etc. Monolithic based approaches basically designed for pedestrian detection while the part-based methods

are designed for a specific body part (like head, shoulder) to estimate the crowd count. Detection based density estimation is successful only in low-density crowd containing the few tens of people in an image. Their performance in highly dense crowd environments is still an open problem. In these environments, usually only partial or some part of the whole objects are visible that possess a great challenge to object detectors for counting or localization. A face is the most visible part of the body, and also detection of crowd faces has further application in face recognition pipelines to identify suspects in mass gathering events (whereas our ultimate goal of crowd analysis is to give a safe and secure environment for public safety). Unfortunately, the crowded scenes are too challenging to detect faces as most of the faces are located far from surveillance cameras or nearby faces are often occluded in the crowd.

Here, we consider the impact of these challenges and present a robust crowd face detection method to estimate the crowd density.

3.1.1 Crowd Detection Method

In this method, the skin color model and Histogram of Oriented Gradient (HOG) features are used for face detection in crowded scenes. HOG offers a robust feature set to differentiate and detect human faces under different illumination conditions, complex backgrounds, a wide variety of poses, etc. The skin color model segregates skin color and non-skin color pixels, which is complimentary to oriented gradient features. Their combination reduces false detection. The RGB and YCbCr color spaces boundary rules are applied to segment skin color regions. The color range is decided by analyzing various images from the existing database. The presented approach involves two phases. In the first phase, the histogram of oriented gradient features is extracted to train Support Vector Machine. In the second phase, skin color segmentation is performed, and oriented gradient features are extracted from the segmented region to classify the crowd faces by the trained model of SVM. The schematic view of our crowd detection approach is depicted in Figure 3.1. Each of these phases will be described in further sections.

3.1.1.1 Skin Color Segmentation

We present a skin color based segmentation technique which is applied to the crowd images to extract foreground area (skin color region) in which the probability of

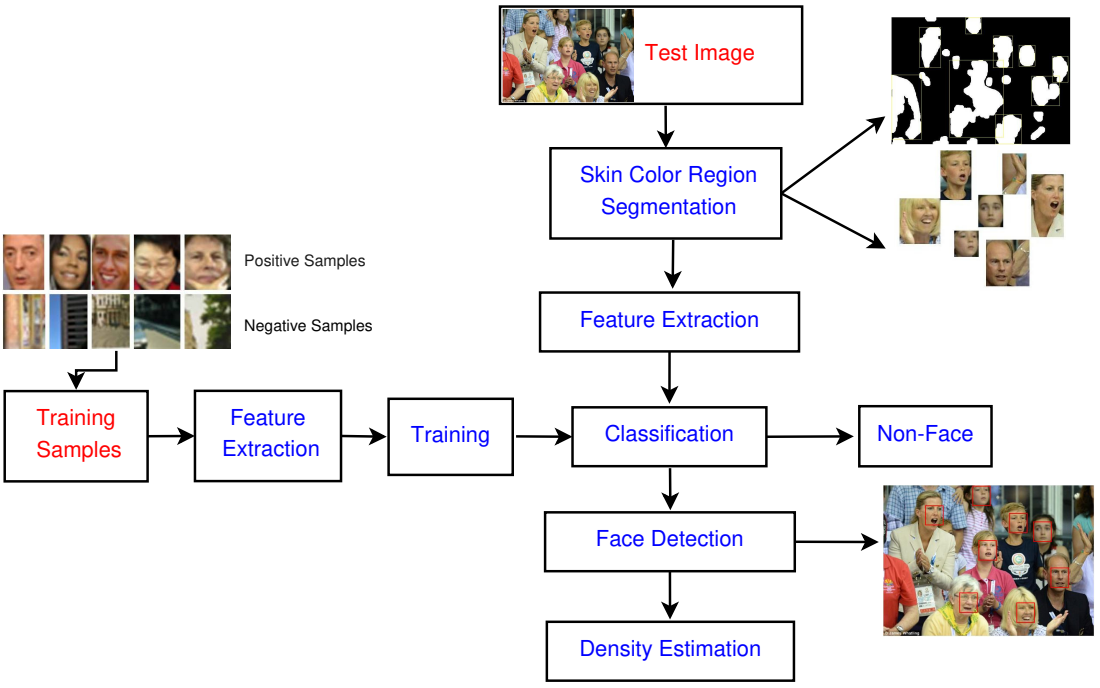

Figure 3.1: Schematic representation of our method for crowd face detection using skin color model and histogram of oriented gradient features.

human face existence is high. The skin color regions are extracted by using the combination of RGB and YCbCr color spaces boundary rules. First, we establish a skin color model that determines the skin tone color range by analyzing thousands of skin color samples of different human races, collected from various sources with different illumination conditions.

In computer vision, many color space models [Vezhnevets et al. (2003)] exist like RGB, HSV, YCbCr, YUV, YIQ, etc. with variable performance. The selection of an adequate color model for skin segmentation is essential as it can prevail the detection rate to a large extent. To improve the detection rate, we have used the combination of RGB with YCbCr because YCbCr provides explicit separation of luminance and chrominance components. The range of skin color value in YCbCr is determined by the skin color distribution graph as shown in Figure 3.2. These color spaces achieve better performance at segmentation and detection. Next, skin color boundary rules are determined for these color spaces. The RGB color space boundary rules are shown in Equation 3.1, for this let the three primary color be equal to W like $W = [R \quad G \quad B]$.

$$\begin{aligned} RGB = R > 95 \vee G > 40 \wedge B > 20 \wedge (\max(W) - \min(R, G < B) \\ > 15 \wedge |R - G| \geq 15 \wedge R > G > B \end{aligned} \tag{3.1}$$

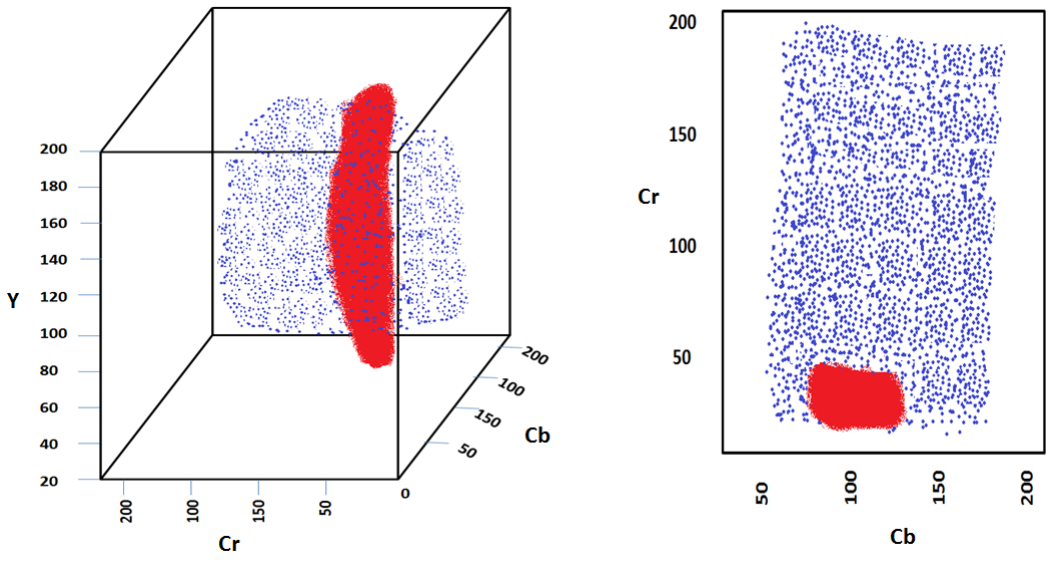

Figure 3.2: (a) Color distribution in YCbCr space, (b) Color distribution in CbCr space (red color represents skin pixel)

$$\text{YCbCr} = (85 \leq Cb \leq 135) \wedge (10 \leq Cr \leq 45) \wedge (Y \geq 80) \tag{3.2}$$

From these boundary rules, we segment the skin color component. Morphological operations are applied to remove tiny objects from the image that are well below the size of a face while preserving the shape and size of larger objects in the image. In our skin color segmentation, these two color models help to increase the face detection rate. The segmentation output of both color models is depicted in Figure 3.3. For each segmented skin color component, the histogram of oriented gradient (HOG) features is computed, which is explained in the following section.

3.1.1.2 Feature Extraction: Histogram of Oriented Gradients(HOG)

Histogram of Oriented Gradients (HOG) [Dalal and Triggs (2005)] is a rotationally invariant feature descriptor that has been used in computer vision, pattern recognition, image processing as well as in optimization problems to detect visual objects. HOG notably outperforms existing feature sets for object detection. HOG aims to generalize an object in such a way that the object produces nearly the same features when viewed in different conditions. These features are computed at the local segment of an image by estimating the occurrence of gradient orientation. HOG is computed on a compact grid of equally spaced cells with overlapping by which detection accuracy is improved. The key advantage of HOG is that it

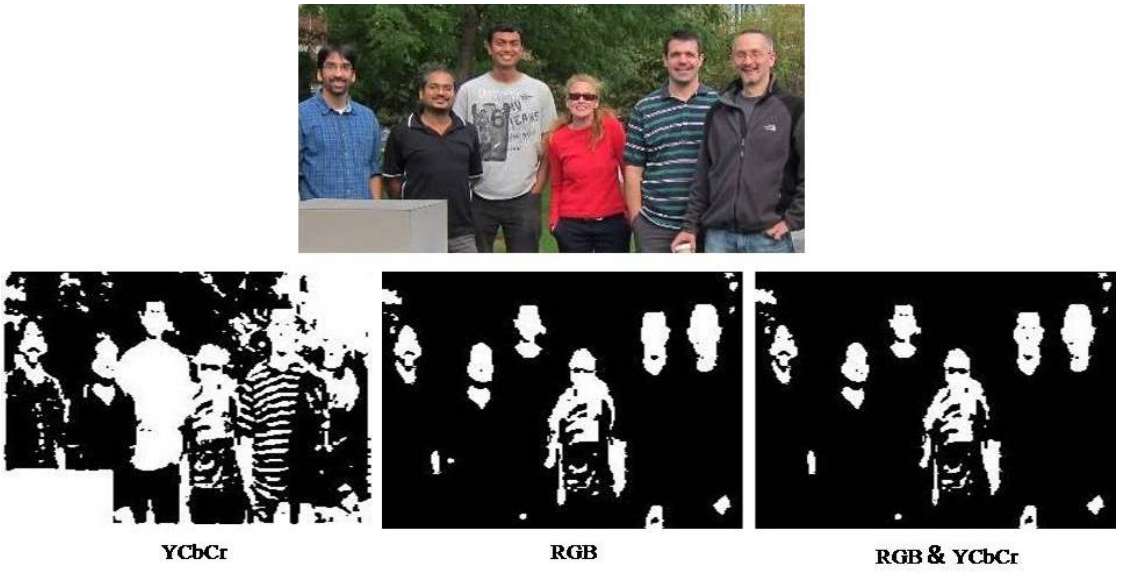

Figure 3.3: Skin color segmentation of an image by using RGB, YCbCr and intersect RGB × YCbCr skin color models.

describes the local shape and appearance of an object just by the distribution of local intensity gradient and edge direction without much knowledge of the gradient position.

The procedure of HOG feature extraction is summarized in Figure 3.5. These features estimate the occurrences of gradient orientation in local parts of a given image. First, the gradient of an image is computed at a dense grid. The image is partitioned into tiny uniformly spaced spatial areas named cells. Next, to form HOG representation, gradient orientations are accumulated for all the pixels of every cell. The cells are normalized using the accumulated local histogram over slightly larger regions called blocks. The normalized features are invariant to illumination or shadowing. These normalized blocks are concatenated to form a feature descriptor. The step by step process is given below to extract oriented gradient features for both training and testing crowd images. The size of each training image is 32×32 pixels.

1. Convolve the image by applying the 1-dimensional centered mask in both horizontal (D_x) and vertical (D_y) directions with the given filter kernels in Equation 3.3. Simple 1-dimensional mask works best [Dalal and Triggs (2005)] as compared to larger masks.

$$D_x = \begin{vmatrix} -1 \\ 0 \\ 1 \end{vmatrix}, D_y = \begin{vmatrix} -1 \\ 0 \\ 1 \end{vmatrix}^T \tag{3.3}$$

Further, subdivide the image into cells wherein every cells is made up of 4×4 pixels and every block is made up of 2×2 cells with 50% overlapping as shown in Figure 3.4.

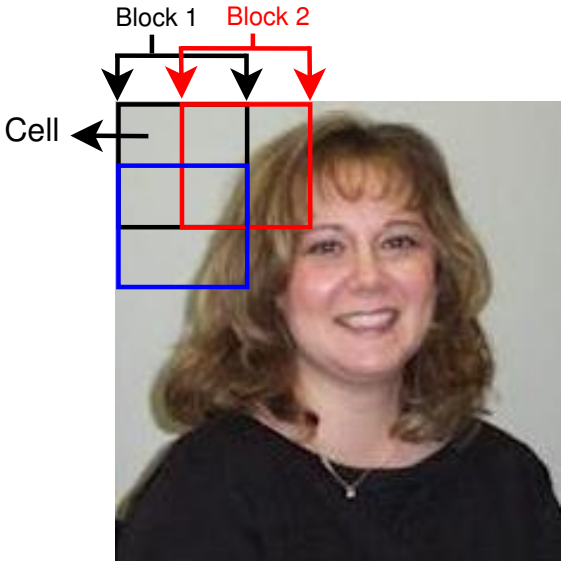

Figure 3.4: Subdivision of image into blocks of 2×2 cells with 50% overlapping and cells with 4×4 pixels.

The selection of cell size depends on image resolution, if the image resolution and the face size is small it is better to use smaller cells as 4×4 or 8×8 but if the resolution is very high you can use larger sizes like $16times16$. For each cell, gradient magnitude(M) and orientation(O) are computed by the following Equations, where i and j are the image(I) pixels.

$$g_x(i,j) = I(i,j-1) - I(i,j+1) \tag{3.4}$$

$$g_y(i,j) = I(i-1,j) - I(i+1,j) \tag{3.5}$$

$$M(i,j) = \sqrt{(g_x(i,j))^2 + (g_y(i,j))^2} \tag{3.6}$$

$$O(i,j) = \tan^{-1}(\frac{g_y}{g_x}) \tag{3.7}$$

2. A cell histogram is computed from the contribution of gradient magnitude. In the case of color image, we opt the channel which has the highest gradient magnitude value for each pixel of an image. The bins of the histogram can

be in a range of 0 to 180 degrees for unsigned gradient and 0 to 360 degrees for a signed gradient.

3. Next, normalize the gradient strength of each cell by combining the cells into larger, spatially connected blocks to make contrast and illumination invariant. The HOG feature vector is a concatenation of all the normalized block regions. These blocks are overlapped together, which means each cell contributes many times into the final HOG feature descriptor. For block normalization, we concatenate all the cell vectors of a block into a larger vector. This vector size should be the number of bins × number of cells in a block. Now, normalization of this feature vector is done by using L2-norm(f), which is computed in Equation 3.8 where v is non-normalized feature vector at each block and e is a small positive constant which averts divisibility of zero in gradient-less blocks. The final HOG feature descriptor is an array of a feature vector of all images which is collected by concatenation of normalized blocks.

$$\text{L2-norm (f)} = \frac{v}{\sqrt{||v||^2+e^2}} \tag{3.8}$$

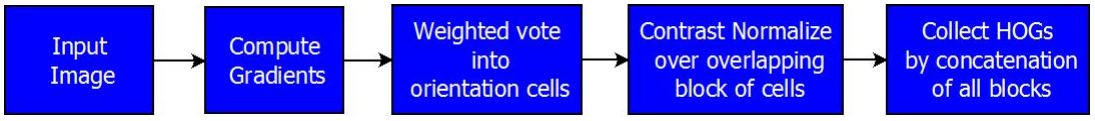

Figure 3.5: Detailed description of HOG feature extraction.

Visualization of HOG features is important to give us the confidence that HOG descriptor is working as it should. The illustrations of the visualization can be seen in Figure 3.6. The extracted features are passed to train the support vector machine.

Figure 3.6: An example: HOG feature visualization of training images used in face detection.

3.1.1.3 SVM Training

Support vector machine [Suykens and Vandewalle (1999)] is the most widely used supervised learning method for classification purposes. The primary task of an SVM is to determine an optimal function of hyper-plane, which classifies or separate the extracted features into a different class. For instance, in Figure 3.7 (a), there can be an infinite linear plane to separate the classes, whereas in Figure 3.7 (b), there is only one plane which maximizes the margin between both class termed as an optimal hyperplane.

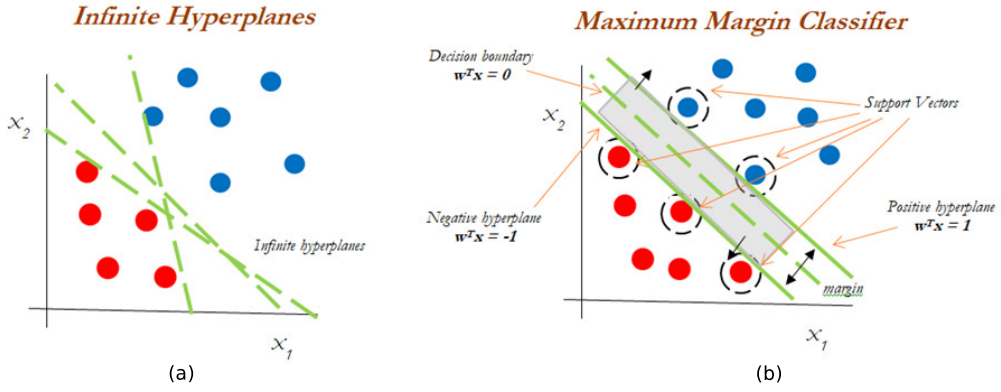

Figure 3.7: Classification between two classes using hyperplane: (a) infinite hyperplanes exist to separate the classes (b) support vector machine determines the optimal hyperplane with maximum margin to separate the classes.

Let suppose, $x_1, x_2, \cdots, x_n$ are the training feature vectors of an input pattern and $y_1, y_2, \cdots, y_n$ are the corresponding labels. Where $x_i \in R^n$, $y_i \in -1, +1$ and $w^T x + b = 0$ is a hyperplane that linearly separates these feature vector based on their label values. The training feature vectors are considered to be optimally separated by the hyperplane if margin of separation is maximal. A separating hyperplane must satisfy the following constraints.

$$y_i \left[(w \cdot x_i) + b\right] \geq 1, i = 1, \cdots, n \tag{3.9}$$

The distance between a feature point x and hyperplane is

$$d\left(w, x\right) = \frac{|w \cdot x + b|}{\|w\|} \tag{3.10}$$

The margin can be defined as $\frac{2}{\|w\|}$ and maximization of margin between classes is equivalent to minimization the Euclidean norm of the weight vector w. Therefore,

the optimal hyperplane is the one which minimizes

$$\Phi(w) = \min \frac{1}{2} \|w\|^2 \tag{3.11}$$

The optimization problem of Equation 3.11 with the conditions of Equation 3.9 is solved by lagrange functional

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i \{y_i [(w \cdot x_i) + b] - 1\} \tag{3.12}$$

where i represents Lagrange multipliers. The Lagrangian function must satisfy the two conditions (i) minimize corresponding to w and b, and (ii) maximize corresponding to $\alpha_i \geq 0$. The primal problem of Equation 3.12 is converted into its dual form because it deals with a convex cost function and linear constraints while the dual one is easier to solve. The dual problem is represented as

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \left\{ \min_{w,b} L(w, b, \alpha) \right\} \tag{3.13}$$

The dual problem can be solved by

$$\bar{\alpha} = arg \min_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{3.14}$$

with constraints,

$$\begin{aligned} \alpha_i \geq 0, i = 1, ...n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{3.15}$$

The solution of Equation 3.13 along with satisfying the constraints of Equation 3.14 and 3.15 computes the lagrange multipliers. The optimal hyperplane is determined by,

$$\bar{w} = \sum_{i=1}^{n} \bar{\alpha}_i y_i x_i \tag{3.16}$$

$$\bar{b} = -\frac{1}{2}\bar{w} \cdot [x_r + x_s] \tag{3.17}$$

where x_r and x_s are support vectors that satisfy the criteria as,

$$\bar{\alpha_r}, \bar{\alpha_s} > 0, \quad y_r = 1, y_s = -1 \tag{3.18}$$

For a new data point x, the classification is done by,

$$f(x) = sign\left(\bar{w} \cdot x + \bar{b}\right) \tag{3.19}$$

For our classification, we only use a linear classifier. Therefore we do not further discuss the non-linear decision classifiers. In our method, the extracted oriented gradients feature vector of all training data along with their corresponding labels is fed to SVM. We perform 5-fold cross-validation over training datasets and generate a trained model. This trained model is utilized to classify the face and non-face in crowd images.

3.1.1.4 Classification

This phase classifies the crowd image into the face and non-face classes. A pre-trained model plus oriented gradients features of each segmented skin color component are supplied to the SVM classifier. SVM classifier categorizes it into a face or non-face. To localize or detect a face, a bounding box is fixed over all the classified faces in an image. Further, we count the classified faces of an image and estimates the density of the crowd in that image.

3.1.2 Experimental Results and Analysis

This section details the experimental setup, datasets, performance analysis metrics used in the analytical analysis of our method and its comparison with some existing state-of-art techniques.

3.1.2.1 Dataset Used for Experiments

We trained the support vector machine by thousands of faces taken from MUCT [Milborrow et al. (2010)] and FEI [Thomaz (2012)] databases which consist of 751 and 2800 face images of different human races, respectively. Various training face images are manually collected and cropped from several datasets to include the wide varieties of faces of different races. Our approach is evaluated on BAO [Frischholz (2012)] and FDDB [Jain and Learned-Miller (2010)] face datasets. We subsequently tested the approach on various manually collected images captured from surveillance areas to determine its accuracy on crowded scenes and to also demonstrate its failure cases.

3.1.2.2 Performance Metrics

The performance evaluation of our approach is done using Miss Rate (MR) and Receiver Operating Characteristic (ROC). Miss rate defines the percentage of the undetected face as defined in Equation 3.20.

$$\text{MR} = \frac{FN}{FN + TP} \tag{3.20}$$

ROC is a probability curve that shows a mapping between the true positive rate (TPR) and false-positive rate (FPR), where TPR and FPR are plotted on y-axis and x-axis respectively. For our evaluations, we model the ROC curve between the TPR and the number of false positives in contrast to the rate of false positives to facilitate comparison with other methods. TPR also termed as sensitivity or recall gives the ratio of detected true positives as compared to the total number of true positives in the ground truth as shown in Equation 3.21

$$\text{TPR} = \frac{TP}{TP + FN} \tag{3.21}$$

Where TP is true positive, i.e., the number of predicted faces which were actually faces, FN is the total number of false negative, i.e., the number of faces which actually belong to the positive class, i.e. face but, have been incorrectly classified as non-faces. FP defines the number of the detected face which are not actually face.

3.1.2.3 Quantitative and Qualitative Evaluation

We evaluated our approach on BAO multi-face and FDDB datasets, which consist of 1500 and 5170 human faces pictured in 157 and 2845 challenging images, respectively.

Evaluation on BAO Dataset

To make a comparison with other state-of-the-art methods on BAO multi-face dataset, we have listed our true positive rate, false positive and miss rate in Table 3.1. This Table demonstrates the comparison of our approach among [Hsu et al. (2002)], Viola Jones approach [Wang (2014)] and skin color approaches [Yadav and Nain (2015, 2016)]. [Wang (2014)] usually fails when eyes are occluded and becomes unreliable when in plane and out of the plane rotation increases more than ± 15 and ± 45 degrees respectively. In [Yadav and Nain (2015)] and [Yadav and Nain (2016)], they used skin color segmentation with the facial feature for face detection. In their approach, visibility of facial feature like eye, nose, lips are required to measure the eccentricity which is used to calculate the probability of a skin color area is a face region. While in our method, proper visibility of mouth, eye, nose, and other facial features are not as important. We trained the SVM with the histogram of oriented gradient feature. It describes the local shape and appearance of an object (face in our case) by the distribution of local intensity gradient and edge direction without much knowledge of gradient position. So a face with a covered mouth, wearing goggles still be detected.

Our algorithm is more efficient as it gives very few false detections (0.72%) as compared to [Hsu et al. (2002)], [Wang (2014)] and [Yadav and Nain (2015, 2016)] methods. In terms of accuracy, we achieve 98.02 value of true positive rate as depicted in Table 3.1, which is higher than the compared methods. The qualitative results of our approach on BAO dataset are depicted in Figure 3.8.

Table 3.1: Quantitative comparison evaluation of our method with other state-of-the-arts mehods on BAO multiple face database

Method	TPR	FP	MR
Viola Jones [Wang (2014)]	82.80	9.5	17.2
Face detection in color images [Hsu et al. (2002)]	89.17	8.1	10.83
Skin segmentation and facial features [Yadav and Nain (2015)]	94.26	6.7	5.74
Our method	98.02	0.72	1.98

Evaluation on FDDB Dataset

Figure 3.8: Qualitative results of the presented approach on sample images of BAO multiface dataset.

We also evaluate our approach on FDDB dataset [Jain and Learned-Miller (2010)] which comprises 5170 faces pictured in 2845 images. The dataset comprises very challenging images of faces such as out-of-plane rotation, convincing occlusion, illumination effects, blur, etc. and some of them have dense crowd environments. [Jain and Learned-Miller (2010)] provide bounding ellipses around faces as ground truth annotations. They also produce cross-folds and binary scripts to validate and evaluate the face detection performance. We evaluate the efficiency of our method by comparing its outcome against ground truth ellipses of an image. The matching score (MS) of our face detection method is computed as in Equation 3.22.

$$MS = \frac{|FD \cap GT|}{|FD \cup GT|} \tag{3.22}$$

where, FD and GT represents an area of detected face and ground truth face respectively in a given image. Every individual detection is associated with two scores, namely continuous and discrete scores. The continuous score is determined by the same matching score as in Equation 3.22. The discrete score acquires a value of 1 or 0 based on the percentage value of the matching score of detection. If the matching score is greater than 50% than discrete score associates a value of 1 otherwise 0 in case of less than or equal to 50%. The true positive rate is defined either as the average continuous or discrete score for the ground truth faces. For both scores, we depict the performance of our method along with state-of-the-art methods by an ROC curve. We compared our approach with various algorithms on FDDB dataset. Most of them have akin characteristics to the Viola-Jones cascade structure. [Mikolajczyk et al. (2004)] adopted local orientation with blob

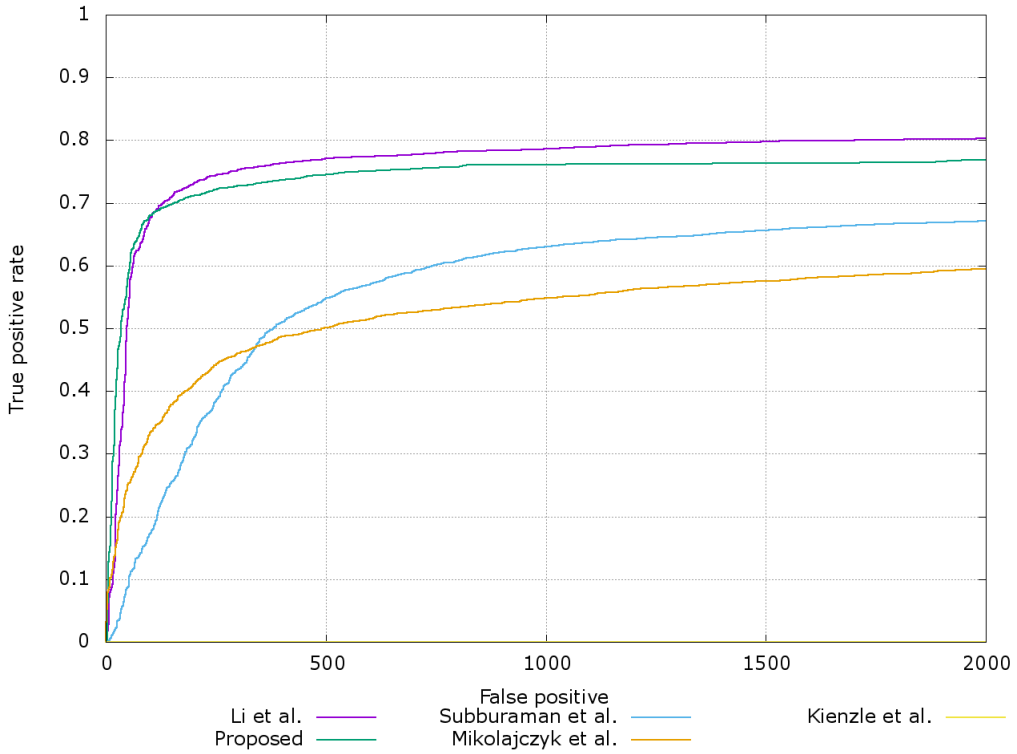

Figure 3.9: Comparison of our approach against the state-of-the-art methods in terms of ROC curve of discrete score on FDDB dataset.

features in a boosted cascade manner to detect faces and other body parts. While [Kienzle et al. (2005)] approximated the Support Vector Machine (SVM) decision function to develop a face detection algorithm and obtain very poor scores like the parallel to x-axis as depicted in both Figures 3.9 and 3.10. Later, [Subburaman and Marcel (2010)] used Modified Census Transform (MCT) features in a boosted cascade and, [Li et al. (2011)] used Speeded Up Robust Features (SURF) in the same boosted cascade architecture and obtained good performance at training and detection time. Our approach outperforms the current state-of-the-art methods except for the SURF based technique, as shown in Figure 3.9 and Figure 3.10 that was introduced by [Li et al. (2011)] on FDDB dataset. Although our approach and [Li et al. (2011)] detect almost equally well, but the bounding boxes generated by SURF detector overlap the faces better than the bounding boxes of our approach.

Evaluation on Manually Collected Images

Our approach is also evaluated on manually collected crowd images from surveillance areas and some Boston Marathon images [Su (2013)] collected from Flickr[1] considering high density, dark illumination, clutter and skin color background.

[1]https://www.flickr.com/photos/tags/crowd/

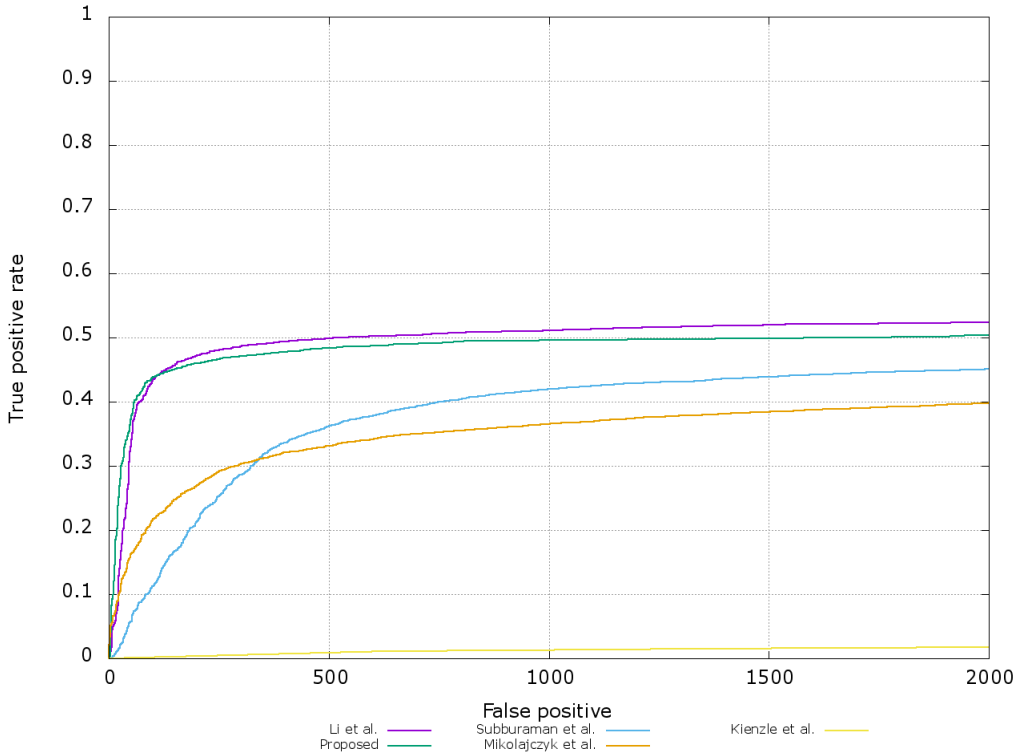

Figure 3.10: Comparison of our approach against the state-of-the-art methods in terms of ROC curve of Continuous score on FDDB dataset.

This database contains 1121 faces pictured in 50 images with an average 20 to 25 people per image. We obtained 78.14% TPR with 21.85% MR as reported in Table 3.2, and qualitative analysis is depicted in Figure 3.11.

We also performed the face detection without skin color segmentation by using HOG only, the true positive rate degrades to 71.36% in the same image sets. It shows the idea of skin color segmentation is very beneficial to improve the correct detection rate.

Table 3.2: Performance evaluation of the presented method on our manually collected images. Some images are of Boston Marathon, and some are collected from mass gathering areas.

Method	Total faces	TPR	FP	MR
HOG	1121	71.36	4.46	28.63
Skin Color + HOG	1121	78.14	4.46	21.85

Figure 3.11: Qualitative results of our approach on manually collected images.

3.1.2.4 Failure Cases

Performance is still far from perfection even though we trained the support vector machine with off-frontal, occluded and blur poses. We have noticed a few different failure cases as shown in Figure 3.12. The miss detections were mostly due to small size, blurry, out of plane and occluded faces. The impact of severe occlusion, blur, clutter background, and small target size should be taken into account since they all frequently occur in the real-world unstructured crowd scene. Our main aim is to deal with a high-density crowd consisting of hundreds or thousands of people per image. In such images, people have severe multiple occlusion, occupying very few pixels per target, very low visibility of head or face, clutter background, perspective effects, etc. Hence, the detection based methods are not feasible for real-world dense crowd scenes. Therefore, we adopted a regression-based approach to analysis high density crowd which will be discussed in upcoming section 3.2.

3.1.3 Summary

We showed that skin color modeling and oriented gradient features are helpful in improving face detection in a crowded scene. The skin color segmentation automatically discards the non-skin objects, in the remaining skin color objects, some false positive may exist due to skin color background, hand, necked neck etc. which are further eliminated by oriented gradient features that provide shape appearances. The presented approach explicitly handles partially occluded or off-frontal poses.

Figure 3.12: Failure cases of face detection in real world crowd scenes, where most of the miss detections are reported due to small size, out of plane and occluded faces.

The gain in robustness resulting from strong training of support vector classifier by a large number of sample from different races under varied illumination and environment. The utilization of this approach is in real-world indoor and outdoor surveillance camera services for crowd density estimation (medium density (up to 50 person per image)). Further, it finds application in face recognition pipelines to identify suspects in mass gathering events.

We compare the approach with current state-of-the-art methods, it is observed that face detection in a high-density crowd in the presence of severe occlusion, blur and perspective effects remains an understudied area. We have shown some failure cases of face detection in the high-density crowd (hundreds or thousands of people) which limits the applicability of detection based methods in such a challenging and high-density crowd images. The inadequacy in robust face detector precludes the use of such detection based approaches for density estimation. Therefore, we adopted a regression-based approach for density estimation in a high-density crowd.

3.2 Regression-based Density Estimation

Regression-based approaches maintain a relationship between local image features to their counts. These methods do not rely on learning detectors as it is relatively a complex task. Regression-based techniques have two major components: (i) low-level feature extraction, and (ii) regression modeling. An image broadly offers two types of features; pixel-based and texture-based features, which are used for encoding low-level information. The pixel-based methods maintain a linear mapping with foreground pixels and human count. [Davies et al. (1995)] and [Hussain et al. (2011)] introduced pixel-based density counting methods that first employ background subtraction and, then carried out a relationship with foreground pixels and density count. [Ma et al. (2004)] also introduced a linear relationship between pixels and people to count human objects. These pixel-based methods cannot be applied in case of severe occlusion in a dense crowd. The texture is considered as an efficient feature to deal with the extremely dense crowd. Various texture-based approaches [Marana et al. (1998, 1999); Xiaohua et al. (2006)] have also been proposed for density estimation in the crowd. In their approaches, the scale of the crowd is limited to few tens to a hundred people, which is not considerable nowadays. The real-world crowd scenes contain an average thousand people due to exponential growth in the worldwide population. It has been observed that dense crowd images depict fine texture, whereas sparse crowd images show coarse texture. With this fact, we present a texture-based multi-source approach, which includes Fourier analysis, Wavelet analysis, SIFT, and HOG features.

3.2.1 Multisource Approach

The main target of our approach is to quantify the number of persons in the given crowd images. The crowd images hardly offer uniformity in terms of density distribution, i.e., every portion of the crowd image has a varying density of people. The diversity in the density distribution of individuals may arise due to the different angle of viewpoint and perspective effects of the camera setup. The whole crowd scene cannot be simultaneously analyzed for density estimation purpose. Therefore, our approach partitions an image into uniformly sampled small patches to make uniform density over the image. The density smoothly varies across the image with the assumption that density should be similar across the adjacent patches. To ensure the continuity of change in density across patches, we apply Markov Random Fields, to handle the issue of smooth variation in density over the

whole image. When we count individuals at cell levels, we consider the uniformity within the cell while assuming in-dependency among the patches. Once the density estimation is performed in each cell, we ignore the independence assumption and put the density count of each cell in multi-scale Markov Random Field to mark the dependency of densities among neighboring patches. The schematic diagram of the presented approach is shown in Figure 3.13, and detailed in upcoming sections.

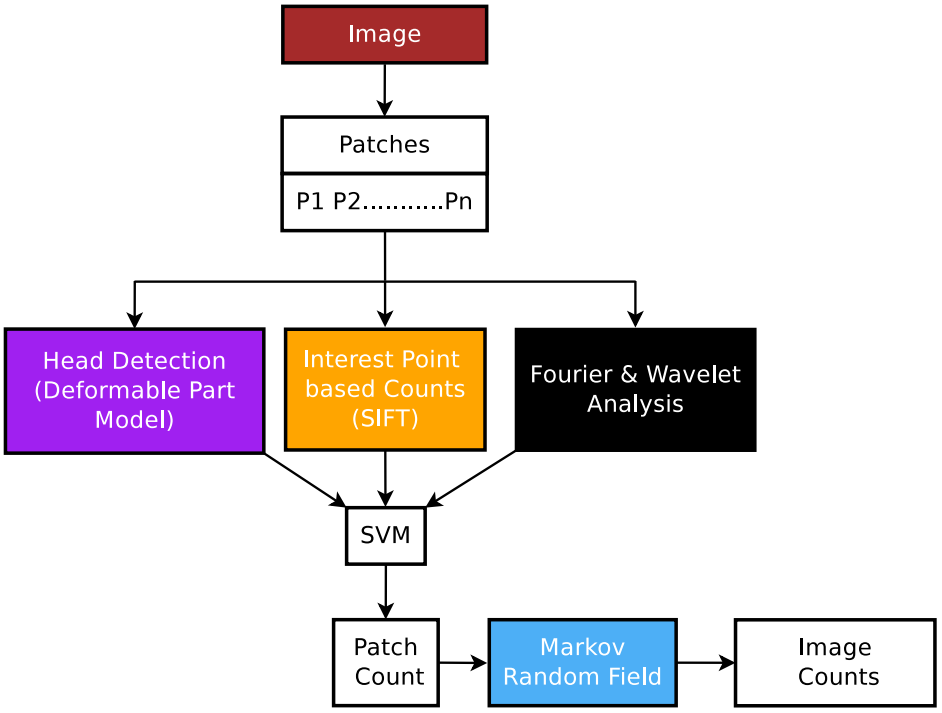

Figure 3.13: A schematic diagram of crowd density estimation approach using multisources of clues via Markov Random Field (MRF).

3.2.2 Patch level Crowd Density Count

We divide every image into small patches, and for each patch, crowd counts and corresponding confidences are estimated from multiple sources as HOG based head detection, Fourier analysis, Wavelet and interest point based (SIFT) techniques. Later, these techniques are combined to obtain final crowd count for that patch using individual counts and confidences.

3.2.2.1 Head-based Count: HOG

Human detection is a challenging task in dense crowd due to severe occlusions. A quick glimpse at crowded scenes announces that the human bodies are almost

entirely occluded, only head is the most visible part at this scale by which we can detect and estimate crowd count. We adopted a Deformable Part Model [Felzenszwalb et al. (2010)] trained on the INRIA Person dataset, and applied only the filter corresponding to head with a much lower threshold. This is because heads are partially occluded and are very small in size in dense crowd images. In feature pyramid $H = (p_0, p_1, \cdots, p_n)$, is an object hypothesis which specifies the location of each head filter in DPM model. The $p_i = (x_i, y_i, l)$ represents the location (x_i, y_i) of filter with level l in feature pyramid (H). The level of pyramid is specified in such a way that the feature vector (f) at that level should be twice the resolution of the root level i.e. $l_i = l_0 - \lambda$ for $i > 0$ where l_0 is root level and λ is level number. In the pyramid H, the feature vector f is present at position p and is represented by $f(H, p)$. The appearance score is defined as the dot product of filter F' and feature vector such as $F' \cdot f(H, p)$. The deformation score of a part is obtained by $d_i \cdot f_d(d_x, d_y)$ where d_i and f_d are deformation cost and deformation features respectively. The confidence score (ConfiScore) as in Equation 3.23 is computed by differencing the summation of the appearance scores of every head filter at their corresponding locations and a deformation score which lies on the location of each portion corresponding to the root with a bias term b.

$$ConfiScore\,(p_0, \cdots, p_n) = \sum_{i=0}^{n} F_i^{'} \cdot f\,(H, p_i) - \sum d_i \cdot f_d\,(dx_i, dy_i) + b \qquad (3.23)$$

where

$$(dx_i, dy_i) = (x_i, y_i) - (2\,(x_0, y_0) + v_0) \qquad (3.24)$$

defines the displacement of i^{th} part and the displacement is correspondent to the anchor position. The following equation defines the deformation features.

$$f_d\,(d_x, d_y) = \left(d_x, d_y, d_x^2, d_y^2\right) \qquad (3.25)$$

Therefore, head detection includes the computation of comprehensive score at each root position according to the best possible placement of the parts.

$$ConfiScore\,(p_{best}) = \max\,(ConfiScore\,(p_0, \cdots, p_n)) \qquad (3.26)$$

The computed comprehensive score of each root location can detect multiple occurrences of a particular object. This approach is equivalent to sliding-window detectors because the $ConfiScore(p_{best})$ offers a detection score for a sliding window, which is described by the root filter. For every detection, the scale and

confidence score is calculated. We may find many misleading head positions in the detection results, as shown in Figure 3.14. However, we find excellent results for nearby heads. HOG based head detection does not impart any useful result in dense crowd images due to fewer pixels per target and invisibility of heads. Therefore, instead of head detection alone, texture-based methods are adopted, which include Fourier analysis, Wavelet transform, and SIFT.

Figure 3.14: Experimental results of head detection: left side image of our dataset provides little bit of considerable outcomes of head detection. The red boxes represent false positives, blue represents false negative, while the yellow one represents correct detections. However, both images are evidence of false negatives and false positives.

3.2.2.2 Texture-based Count

The texture of dense crowd images is repetitive in nature. When we look at dense crowd images, all people appear almost similar from a certain distance due to recurrent crowd texture. Here, we employ two texture-based approaches: Fourier and wavelet analysis. They give per patch count, which will be used posterior to ensure the final image count.

Fourier Analysis

In crowd images, when thousands of people are present, each people occupy only a few pixels per target, and some of them are also very distant from the camera viewpoint. In such scenes, the histogram of oriented gradient does not provide any beneficial information. A dense crowd is immanently monotonous in nature due to the same appearance from a distance. The density of the crowd at the patch level is uniform due to the repetitive texture pattern. Therefore, crowd texture can be captured by Fourier Transform (FT). In the frequency domain, the sporadic

occurrence of peaks is considered as human heads, as shown in Figure 3.15. Particularly, for a given patch (P), we convert that patch into gradient image patch $\partial(P)$ followed by low pass filter to ignore high-frequency component. Next, low amplitude frequencies are discarded, and inverse Fourier Transform (FT') is applied for reconstruction of image patch(P_r). After applying non-maximal suppression, we select the local maxima in the reconstructed image patch. The number of local maxima as in Figure 3.15 is defined as a Fourier count estimation, which is denoted by red dots. Additionally, we calculate various statistical parameters such as entropy, mean, variance, skewness, and kurtosis for both reconstructed and difference ($|P_r - \partial(P)|$) image patch. The Fourier count and these computed parameters are further used as input for regression modeling.

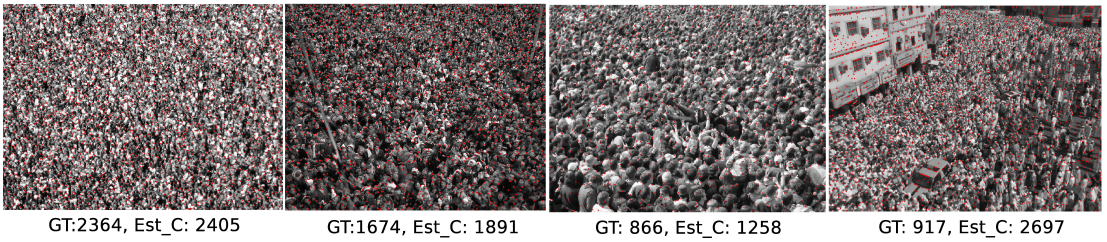

Figure 3.15: The local maxima (red points) considered as head peaks are obtained by applying the inverse Fourier transform. We observe that in first three images, the local maxima peaks corresponding to head are very well located in a dense crowd, while the fourth image is crowd blind and can not count the approximate presence of the crowd.

Wavelet Analysis

The wavelet approach [Chang and Kuo (1993)] for texture analysis have been widely and successfully employed in some real world applications. It is a very powerful model for texture discrimination and has the ability to decompose complex information and patterns into elementary forms. Wavelet transform can also be used to reconstruct the 3D surface geometry from 2D image (i.e. shape from texture), which is a major issue in texture analysis. The wavelet transform decomposes a signal into the family of functions as:

$$\psi_{m,n}(x) = 2^{\frac{-m}{2}} \psi[2^{-m}(x-n)] \tag{3.27}$$

generated by dilation and translation operations by using a prototype function ψ. The prototype function has to satisfy the below criteria:

$$\int \psi(x)\, dx = 0 \tag{3.28}$$

The computation of a wavelet transformation of an image $f(x)$ can be seen by analysis and synthesis equations as given below:

$$C_{m,n} = \int_{-\infty}^{+\infty} f(x)\,\psi_{m,n}(x)\,dx \quad where \quad f(x) = \sum_{m,n} C_{m,n}\psi_{m,n}(x) \tag{3.29}$$

The below scaling function $\phi(x)$ formulates the mother wavelet

$$\phi(x) = \sqrt{2}\sum h_0(k)\,\phi(2x-k) \quad \psi(x) = \sqrt{2}\sum h_1(k)\,\phi(2x-k) \tag{3.30}$$

where $h_1(k) = (-1)^k h_0(1-k)$ The wavelet model can be generalized to any dimension. We use a two dimensional wavelet transform with multi-resolution properties to extract texture features for crowd counting framework. The basic function of two dimensional wavelet transform the separable product of scaling function ψ and ϕ expressed as:

$$\psi_1(x,y) = \phi(x)\psi(y), \quad \psi_2(x,y) = \psi(x)\phi(y), \quad \psi_3(x,y) = \psi(x)\psi(y) \tag{3.31}$$

The multi-resolution property of 2D-wavelet intends to transform an image in such a representation which contains both spatial and frequency information. For a given patch P, a three-step pyramid-structured wavelet transformation is employed to achieve the 10 lower resolution sub-images. Each sub-image offers energy calculated as:

$$e_{2^j} = \frac{1}{M \times N}\sum_{x=1}^{M}\sum_{y=1}^{N} I_{2^j}^2(x,y) \tag{3.32}$$

where $I_{2^j}^2(x,y)$ represent any of the sub-images with $M \times N$ resolution, and the calculated energy represents the extracted feature vector.

Thus, corresponding to every patch of an image, a 10-dimensional feature vector is obtained. Different texture patterns of an image generate different feature vectors. Hence, the above-calculated energy feature is used for discriminating crowd and non-crowd regions. These energy feature vectors and corresponding ground truth counts of training images are used to train a support vector regression. The previous statistical parameters viz. variance, skewness, and kurtosis of the 10 lower resolution sub-images are also calculated here and passed to the regression phase along with the patch count.

3.2.2.3 Interest Points based Count: SIFT

We use the quantized SIFT features used by [Idrees et al. (2013)] to divide an image into positive and negative density regions. [Idrees et al. (2013)] focused on interest point based SIFT descriptor for crowd counting and to obtain a confidence score of an image patch whether the corresponding patch contains crowd or not. We observe an interesting fact that sky, walls, and trees usually appear in outdoor scenes where head detection results in a high rate of false positives. In such outdoor scenes, the frequency based Fourier Analysis is unable to discriminant the crowd texture. Therefore, it is necessary to ignore counts from such regions of an image. Hence, we used interest point based SIFT descriptor to estimate crowd counts and obtain confidence score of the presence of the crowd. We extract SIFT features by utilizing VL-FEAT [Vedaldi and Fulkerson (2010)] library from training images and cluster the interested key-points in a codebook of size S. We train a regression-based support vector with sparse SIFT features and ground truth count of each patch. The trained SVR model is used to obtain counts for testing image patches.

The probability of observing crowds in a specific image patch can be modeled as spatial Poisson Counting Process (PCP) with λ density. The PCP is $N(P) \sim Poisson(\lambda\,|P|)$ i.e. the expected value of $N(P)$ is equivalent to $\lambda\,|P|$. For the image I, total density is the summation of the density of all patches. Since we have assumed that the density among patches is independent and homogeneous in a patch.

$$N(I) = N(P_1 \cup P_2 \cup \cdots P_n) = N(P_1) + N(P_2) + \cdots + N(P_n) \tag{3.33}$$

where $P_1, P_2, ..., P_n$ are disjoint patches of an image I. In a given patch, the recurrence α of a specific feature i can be modeled by Poisson Random Variable. The probability of a patch containing crowd with expected density $\lambda^{+}$is defined by

$$P(\alpha_i|crowd) = \frac{\exp(-\lambda_i^{+}) \cdot (\lambda_i^{+})^{\alpha_i}}{\alpha_i!} \tag{3.34}$$

Then as well, for a patch with no crowd with expected density λ^{-}:

$$P(\alpha_i|{-}crowd) = \frac{\exp(-\lambda_i^{-}) \cdot (\lambda_i^{-})^{\alpha_i}}{\alpha_i!} \tag{3.35}$$

Therefore, the relative density of a particular feature will differ in the crowd and non-crowd patches and can be used to locate crowded and non-crowded regions in

an image. By our independence assumption of counts among patches, the count of any two SIFT words in a patch is:

$$P(S_i, S_j|crowd) = P(S_i|crowd)P(S_j|crowd) \tag{3.36}$$

The crowd and non-crowd patches maintain the log-likelihood ratio $\sigma(P)$ which is interpreted as the confidence of detecting crowd in an image patch.

$$\begin{aligned} \log(\alpha_1, \alpha_2 ... \alpha_S \mid crowd) - \log(\alpha_1, \alpha_2 ... \alpha_S \mid \alpha crowd) = \\ \sum \left(\lambda_i^- - \lambda_i^+ + \alpha_i \left(\log \lambda_i^+ - \log \lambda_i^-\right)\right) \end{aligned} \tag{3.37}$$

3.2.3 Synthesizing Multiple Complementary Sources: SVR

To learn and synthesize at the patch level, the patches are sampled from the training images of the dataset. The multi-source methods explained above provide a feature vector along with confidence score and other statistical features. Support Vector Regressor (SVR) is trained by using the ground truth annotations and computed features, which aggregates the information obtained from multi-sources to generate a patch count estimation. The total crowd density count in an image is achieved by aggregating the patch counts of corresponding image patches in the grid structure. Initially, we assume neighboring patches are independent. Later, we remove the independence assumption and mark the dependency of counts among neighboring patches by putting every patch count in multi-scale Markov Random Field.

3.2.4 Consistency Constraint: Markov Random Field

To remove discontinuity and impose smoothing among density counts at the patch level, all patch counts are placed in Markov Random Field. Furthermore, density around small patches maintains consistency with fewer texture repetitions. In large patches, if the density is consistent, it results in more people, more repetitions, and prominent significant frequency proportion. Therefore, it is crucial to a priori determine the accurate proportion for an image analysis, which extends the problem to a multi-layer MRF. The notations for the problem of discontinuities among patch count can be defined as follows: in image grid graph, P and L are the set of patches and their corresponding labels of an image, respectively. The labels

refer to the values on which we want to impose smoothness at each patch. A labeling l accredits a label $l \in L = [0, 1, 2, \cdots, C_{\max}]$ to each patch $p \in P$. We consider that the value of label differ smoothly among patches but may change drastically at boundary lines. In multi-scale MRF, the label characteristic is defined by an energy function as:

$$E(l) = \sum_{p \in P} D_p(l_p) + \sum_{(p,q) \in N} V(l_p - l_q) \tag{3.38}$$

where N represents four-connected neighbors at same level. The label assigning cost to the two neighboring patches l_{p1} and l_{p2} is defined by $V(l_p - l_q) = \min\left((l_p - l_q)^2, \tau\right)$ which is also known as discontinuity cost or smoothness term. Data cost $D_p(l_p) = \lambda(\eta_p - l_p)^2$ refers to the cost of assigning label l_p to a particular patch p which is calculated independently from the patches at layers above and below it. The Max-Product/min-sum BP algorithms [Felzenszwalb and Huttenlocher (2006)] are used to calculate the minimum cost labeling of energy function on image grid structure. The Max-Product algorithm details the probability distributions but an equivalent computation can deal with negative logarithm probabilities, so the max-product converts into min-sum algorithm. This algorithm processes by sending message among the patches by using 4 connected component of an image. A message is a sequence of labels stored in a vector form. Let at any particular time t, a patch node p_1 sends a message $m^t_{p1 \to p2(l_{p2})}$ to neighboring patch node p_2 computed by:

$$m^t_{p1 \to p2(l_{p2})} = \min_{l_{p2}} \left(l_{p1} - l_{p2} + D_{p1}(l_{p1}) + \sum_{s \in N(p1) \setminus p2} m^{t-1}_{s \to p1}(l_{p1}) \right) \tag{3.39}$$

where $N(p1) \setminus p2$ refers the neighbors of $p1$ apart from $p2$. Further, for an image patch $p2$ with label l_{p2}, a belief vector is obtained as:

$$b^t_{p2}(l_{p2}) = D_{p2}(l_{p2}) + \sum_{p1 \in N_{p2}} m^t_{p1 \to p2}(l_{p2}) \tag{3.40}$$

The interpretation follows by spanning Equation 3.39 with four neighbors at the lowest layer. The belief vector for each patch is computed by Equation 3.40. Next, the beliefs of every 2×2 patch are summed up, providing the beliefs for the next nodes b^t_i till the one above the lowest layer. After finishing the four spans at

the middle layer, the next span of message transfers to the middle layer from the intermediate patch, then the belief computation is done at the middle layer. The same steps are repeated for the top layer, and the entire process completes in a single time sequence of t. Moreover, the entire steps are repeated from lowest to top layer. The beliefs of an intermediate patch are distributed among every down patch, i.e., each patch q in 2×2 group always shares the beliefs from the above layers which are depicted by Equation 3.41

$$b_{i,q}^{t+1}(l_q) = b_q^t(l_q) \cdot b_i^{t+1}(l_q) \setminus b_i^t(l_q) \tag{3.41}$$

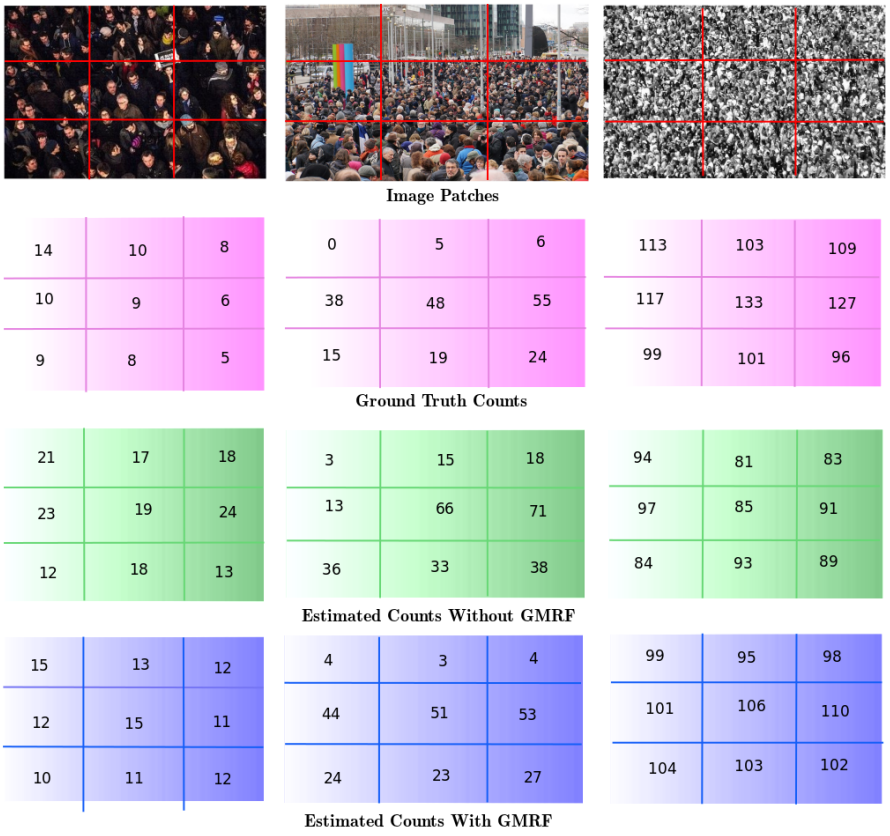

Figure 3.16: Results after imposing GMRF at one layer: The first row shows three patches from the random images of the dataset. The corresponding ground truth counts are shown in the second row, and the third and fourth rows depict the estimated count without and after applying GMRF respectively.

Subsequently, the final beliefs are computed by using Equation 3.41 after a certain number of reiterations. We choose the final labels (counts) with a minimum cost from the belief vectors. The lowest layer offers the image count in terms of the summation of labels at that layer. The patch count between adjacent layers can be calculated simultaneously due to independent data term at each layer. Three patches are shown in Figure 3.16, depicting the improved estimated count based on

MRF. In these instances, the overestimated count is reduced near to ground truth values after imposing a smoothing consistency constraint. A particular example is considered in the middle column. The patch with GT = 38 calculated much lower count compared to its neighbor, but after imposing the smoothing constraint, the new count is closer to its neighbors.

3.2.5 Experimental Results and Analysis

This section states different measures used for performance evaluation and discusses the results of the presented method by showing quantitative analysis of the various datasets. We also compared our method with existing state-of-art techniques and analyzed the results. The combined comparative analysis of our approach with other existing techniques of crowd density estimation on different datasets is reported in Table 3.4.

3.2.5.1 Performance Measures

We adopted two standard measures for performance evaluations which are Mean Absolute Error (MAE) and Mean Squared Error (MSE) as defined in Equation 3.42. The MAE defines the accuracy of density count while the MSE defines robustness of count.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - \hat{z}_i| \quad and \quad MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\{z_i - \hat{z}_i\}^2} \tag{3.42}$$

Where, N represents the number of test images or patches, $\hat{z}_i$ represents the number of the estimated count and z_i is the ground truth count corresponding to i^{th} sample. Since we are dealing with patches, the presented approach is evaluated at both image and patch level.

3.2.5.2 Datasets Used for Experiments

We evaluated our presented framework on four different datasets which include publicly available UCSD pedestrian dataset [Chan et al. (2008)], Shanghai Tech_A [Zhang et al. (2016)] dataset, UCF_CC_50 [Idrees et al. (2013)] dataset and extended UCF_CC_100 dataset. The UCF_CC_50 dataset was missing the images

Table 3.3: Summarization of statistics of four datasets, where Max, Min, and Avg represent the maximum number of people, minimum number of people and average count in images respectively.

Dataset	**No of Images**	**Min**	**Max**	**Avg**	**Total**	**Resolution**
UCF_CC_50	50	94	4543	1279	63,974	Varied
Extended UCF_CC_100	100	81	4633	871	87,135	Varied
Shanghai Tech Part A	482	33	3139	501	241,677	Varied
UCSD	2000	11	46	25	49,885	158*238

with a wider variety of perspective distortions. Thus, we augment the UCF_CC_50 dataset by accumulating 50 new images collected by [Bansal and Venkatesh (2015)] to ensure the robustness of the presented method. The images of the extended UCF_CC_100 dataset have widely varying viewpoints with severe multiple occlusions. The descriptions of all these datasets are summarized in Table 3.3, and some sample images are depicted in Figure 3.17.

3.2.5.3 Quantitative Evaluation

We quantitatively evaluate our approach on four different datasets namely UCF_CC_-50, UCF_CC_100, UCSD and Shanghai Tech_A datasets. The detailed description of each dataset is given in the following sections.

(a) UCF_CC_50 Dataset

Our first evaluation is performed on UCF_CC_50 dataset [Idrees et al. (2013)] which is publicly available. This dataset consists of 50 images ranging from 96 to 4633 persons. The authors of [Idrees et al. (2013)] provided the annotated ground truth details for every image of this dataset. There are total 63974 annotations of human heads in the 50 images. We partitioned the dataset into 5 groups, each having 10 images, i.e. 5-fold cross-validation. The comparative results of our approach with the existing methods [Idrees et al. (2013); Lempitsky and Zisserman (2010); Rodriguez et al. (2011a)] are reported in Table 3.5. Being independent of

Figure 3.17: Illustrations of sample images from four datasets: (a) UCF_CC_50 [Idrees et al. (2013)], (b) UCSD [Chan et al. (2008)], (c) Shanghai Tech_A [Zhang et al. (2016)], (d) UCF_CC_100 [Bansal and Venkatesh (2015)]. The datasets (a), (c) and (d) have dense crowd images with varied scenes while the dataset (b) has relatively very low density images with no variation in perspective across images.

Table 3.4: The combined comparative results of our approach with existing methods for crowd density estimation on UCF_CC_50, Shanghai Tech_A and UCSD datasets.

Approaches	**Datasets & Performance Measures**					
	UCF_CC_50		**Shanghai Tech_A**		**UCSD**	
	MAE	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
Rodriguez et al. (2011a)	655.7	697.8	–	–	–	–
Lempitsky and Zisserman (2010)	493.4	487.1	–	–	–	–
Idrees et al. (2013)	419.5	541.6	–	–	–	–
Zhang et al. (2015)	467.0	498.5	181.8	277.7	1.60	3.31
Zhang et al. (2016)	377.6	509.1	110.2	173.2	1.07	1.35
Kang et al. (2018)	406.2	404.0	–	–	1.12	2.00
Chen et al. (2012)	–	–	–	–	2.25	7.82
Chan et al. (2008)	–	–	–	–	2.24	7.97
Marsden et al. (2016)	–	–	126.5	173.5	–	–
Our approach	376.1	477.8	123.6	167.3	1.09	1.25

videos and detection techniques, these methods are best suitable for comparative analysis. [Rodriguez et al. (2011a)] used head detection based approach for density estimation while [Lempitsky and Zisserman (2010)] adopted SIFT features with regression function to crowd count. The authors of [Idrees et al. (2013)] claimed about the performance of [Rodriguez et al. (2011a)], which is best suited for counts of approx 1000, but above 1000 counts, there is an increase in error due to the dependency of the dataset, it does not work well for varying crowd density. The errors increase dramatically for both very low and very high-density levels. However, [Lempitsky and Zisserman (2010)] offer good accuracy at higher density levels, but the performance degrades at lower density levels. Our method outperforms all the compared methods and performs quite well at both medium to high-density levels ranging greater than 1000 individuals per image.

Table 3.5: Comparative results of our approach with the methods of [Rodriguez et al. (2011a)], [Lempitsky and Zisserman (2010)], and [Idrees et al. (2013)] in terms of MAE and MSE on UCF_CC_50 dataset.

Methods	**MAE**	**MSE**
Rodriguez et al. (2011a)	655.7	697.8
Lempitsky and Zisserman (2010)	493.4	487.1
Idrees et al. (2013)	419.5	541.6
Our approach	376.1	477.8

The group level qualitative analysis of our approach with the other methods is depicted in Figure 3.18. The UCF_CC_50 dataset is divided into ten groups of five images, and the images are placed in increasing order of the ground truth counts. The x-axis represents each of the 10 groups, and the y-axis represents total average counts estimated by each of the compared approaches. We observe that all the compared methods including presented one, underestimate the tenth set when the density increases beyond 2200, and this can be due to histogram features that capture relative texture frequencies. The relative frequency is hardly distinguishable in very high-density images.

(b) UCF_CC_100 Dataset

The second experiment is done on the augmented UCF_CC_100 dataset, which contains 100 images. As compared to UCF_CC_50, the UCF_CC_100 dataset have much more diverse images having varying crowd densities at different viewpoints. These enriched features of this dataset help in evaluating the robustness of the presented method more efficiently. In this experiment, the dataset is divided into

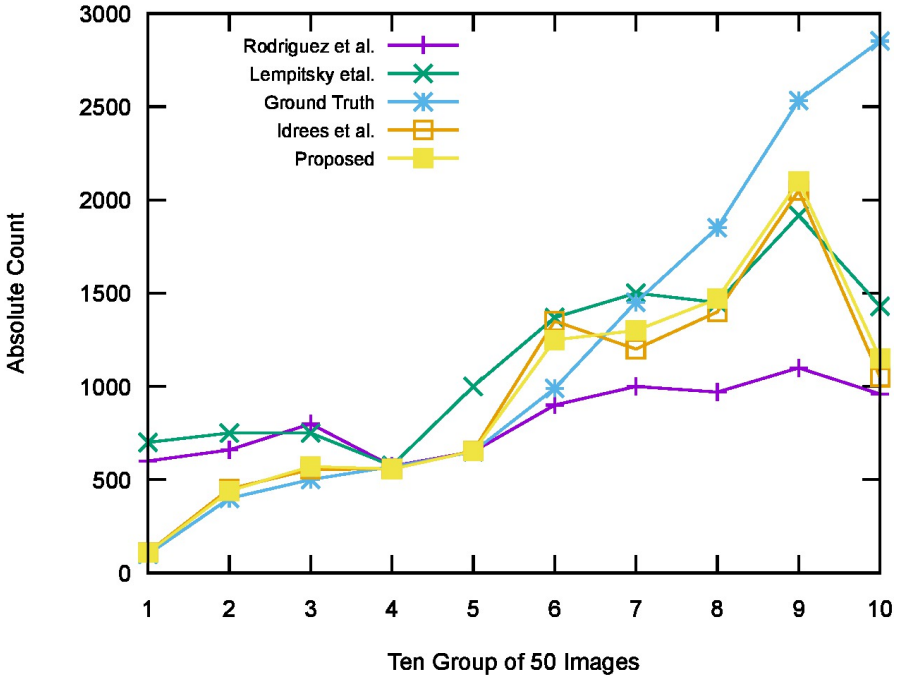

Figure 3.18: The group level qualitative analysis of our approach with the other methods [Rodriguez et al. (2011a)], [Lempitsky and Zisserman (2010)], and [Idrees et al. (2013)] on UCF_CC_50 dataset. The comparison is done between actual ground truth and estimated counts given by different methods. All the methods underestimate the density in the tenth group where the density is > 2200.

groups of 25 images with 4-fold cross-validation. For quantitative analysis, the influence of each individual and the complementary source is shown in the second column of Table 3.6. The results of Fourier Analysis count (first row of the second column) providing MAE and MSE of 803.9 and 782.1, respectively. Supplementing the result in wavelet analysis reduces the mean absolute error by 292.7. Adding the results from head detections improves the MAE to 507.2 (a little bit). With the inclusion of interest point SIFT features, both performance measures reduce to 498.6 and 568.6. Eventually, by imposing smoothness in order to maintain consistency through MRF, both the measures improve the performance by reducing MAE and MSE to 489.3 and 557.1, respectively. Table 3.6 demonstrates that the error measures are reducing as we add more complementary sources, and the MSE follows the same trend as MAE, which defines exactness and completeness, respectively.

We also analyze and demonstrate the result of each image in terms of estimated count and ground truths. In Figure 3.19, the blue dots represent the estimated counts per image, and orange diamonds are the exact ground truth of images where the images are sorted with respect to actual ground truths. The actual error per patch is very small or almost constant until 85 images (sorted). We analyze that

Table 3.6: Quantitative results analysis of each individual and complementary sources are quantified in terms of MAE and MSE on extended UCF_CC_100 and Shanghai Tech_A datasets. The result claims that each source is complementary to others and improves performance.

Presented Methods	**Datasets & Performance measures**			
	Shanghai Tech_A		**Extended UCF_CC_100**	
	MAE	**MSE**	**MAE**	**MSE**
Fourier analysis (F)	385.6	406.1	803.9	782.1
F + Wavelet (W)	261.0	297.9	511.2	612.1
F + W+ HOG head (H)	250.4	295.6	507.2	612.1
F + W + H+ SIFT (S)	152.3	203.0	498.6	568.6
F + W+ H + S + MRF	123.6	167.3	489.3	557.1

the presented approach produces a high error for only those images which have a very high density of people where crowd density is > 2500. One of the reasons behind poor density count is, some of the images in the dataset suffer from lens distortion, which leads to the high error rate.

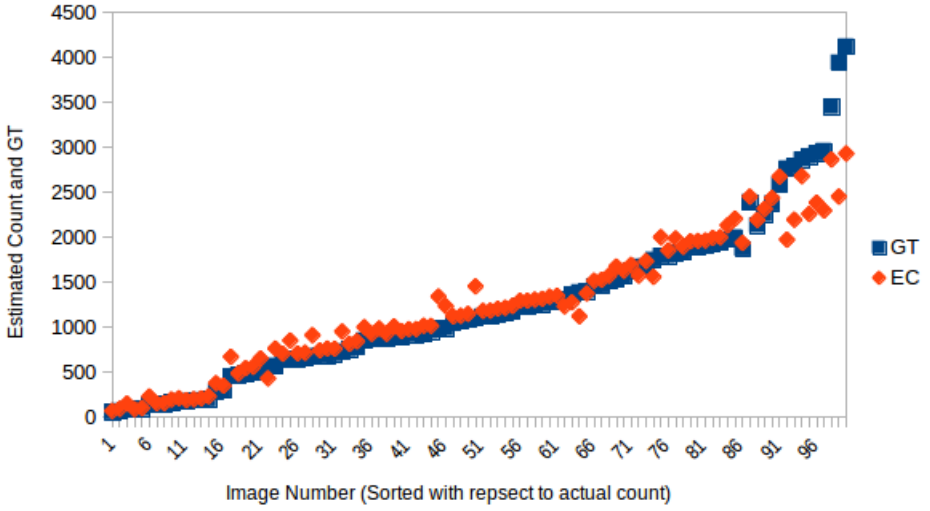

Figure 3.19: Per image error analysis in terms of actual error count. The error is the difference of estimated count (EC) and ground truth (GT). The total estimated count of an image is the summation of patches of that image. The X-axis represents image number, which is sorted (increasing order) regarding actual density count. The Y-axis represents ground truth and estimated count in terms of blue and orange dots respectively.

(c) UCSD Dataset

We also test the presented approach for a single scene UCSD dataset [Chan et al. (2008)], which contains 2000 images. Although this is not a suitable dataset for our method as we are focused on the high-density crowd. But to measure the efficacy and robustness of the algorithm, we consider this dataset and pursue the same dataset settings as in [Chan et al. (2008)]. The dataset is partitioned into training and testing images. The images from 601 to 1400 are considered as the training sets, and the rest 1200 images are treated as testing sets. Each image is divided into patches of size 72×72. The training patches are randomly extracted but to test the dataset, a sliding window approach is used without overlapping. Table 3.7 reports the comparative performance of our algorithm with two methods [Chen et al. (2012)] and [Chan et al. (2008)] which are based on regression. It is important to mark that our method is independent of any foreground information, while the compared methods are dependent on the foreground segmentation features. Our method outperforms these regression-based approaches for both MAE and MSE error metrics. As per Table 3.7, error values of our approach significantly lower than the compared methods. This demonstrates the strength of our method, which can estimate the density count in both extremely dense and sparse crowd images.

Table 3.7: Comparative results of our approach with the methods of [Chen et al. (2012)] and [Chan et al. (2008)] on the UCSD crowd counting dataset.

Approach	MAE	MSE
Ridge Regression [Chen et al. (2012)]	2.25	7.82
Gaussian Process Regression [Chan et al. (2008)]	2.24	7.97
Our approach	1.09	1.25

(d) Shanghai Tech_A Dataset

We also evaluate our approach on recently introduced Shanghai Tech_A dataset [Zhang et al. (2016)] which contains 482 images with $241,677$ head annotations. We arbitrarily divide the dataset into ten groups, with each group comprising 48 images. Therefore, we execute the 10-fold cross-validation for testing the performance. The influence of complementary sources is reported in the second column of Table 3.6. For this dataset, our algorithm is compared with the work of [Zhang et al. (2015, 2016)] and [Marsden et al. (2016)] which use a convolutional neural network (CNN) for crowd density estimation. The authors of [Zhang et al. (2016)] also use the Local Binary Pattern and perform ridge regression to estimate the

image count. It is observed that the single source based result is rather weak. The comparative analysis of our approach with all these methods is depicted in Table 3.8. For the performance metric MAE and MSE, we outperform the method [Zhang et al. (2015)] and [Marsden et al. (2016)], and beaten by only [Zhang et al. (2016)] in terms of MAE which is a much more computationally expensive CNN based model. All CNN based methods require upscale GPUs and the massive amount of datasets, while the presented method works very well in a cost-effective manner.

Table 3.8: Comparative performance analysis of our method with convolutional neural networks (CNN) based methods on Shanghai Tech_A dataset in terms of MAE and MSE. The presented approach outperforms [Zhang et al. (2015)] and [Marsden et al. (2016)] and is only outperformed by a computationally expensive method [Zhang et al. (2016)] in terms of MAE.

Method	**MAE**	**MSE**
Zhang et al. (2016)	303.2	371.0
Zhang et al. (2015)	181.8	277.7
Marsden et al. (2016)	126.5	173.5
Zhang et al. (2016)	110.2	173.2
Our approach	123.6	167.3

3.2.6 Summary

The presented approach estimates the density of people in exceptionally dense crowd images. We employ a texture-based multi-source technique that grasps the multiple information parameters gathered from various sources such as Fourier, Wavelet, head and SIFT in terms of confidence scores and different statistical parameters. Every source provides a separate estimate of density count at patch level, which is further integrated to provide a total count of an image. Then, a Markov Random Field is applied to improve estimate over the image by ensuring consistency and smoothness on nearby patches.

The presented approach performs well at various densities and gives persistent error value over the entire images with variable crowd count. We report our experiments on four datasets: UCF_CC_50, UCF_CC_100, Shanghai Tech_A, and UCSD. The former three are complementary to the presented approach as they contain incredibly dense crowd images of real-world applications, and later one tests the robustness of our approach by performing well in low density too. The experimental results achieve remarkable performance in terms of MAE and MSE when compared with existing methods of crowd density estimation. Also, the pre-

sented framework is computationally simple as compared to the recent well-known deep learned CNN. Being computationally light-weighted, it can be applied to real-time density estimation in crowded areas, which presents a danger of stampede and can provide guidelines for evacuation strategies in the emergency.

3.3 Chapter Summary

This chapter presents detection and regression-based approaches for density estimation in exceptionally dense crowd images. The crowd detection approach incorporates a skin color model, oriented gradient features, and support vector machine. The SVM strongly trained by a large number of face samples of different races under varied illumination and environment conditions. The skin color modeling, along with oriented gradient features, helps to reduce false-positive results. The presented approach explicitly handles partially occluded or off-frontal poses and obtain good results at nearby or larger faces in medium-density crowd.

We compared the approach with current state-of-the-art methods, and it is observed that face detection in a high-density crowd in the presence of severe occlusion, blur, and perspective effects remains an understudied area. We have shown some failure cases of face detection in the high-density crowd (hundreds or thousands of people), which limits the applicability of detection based methods in such a challenging and high-density crowd images. The inadequacy in a robust face detector precludes the use of such detection based approaches for density estimation. Therefore, we adopt a regression-based approach for density estimation in the high-density crowd.

In a regression-based approach, we combine counts and confidences obtained from four different sources and then impose consistency constraints in neighborhood patches to revise the count of incorrect patches, thereby better estimates are produced for the entire image. Experimental evaluation of both approaches depicts good confidence for their efficacy in density estimation of dense crowd images. The next chapter discusses an automatic analysis of crowd scenes. For this purpose, we describe our crowd flow segmentation approach in high-density crowd videos.

Chapter 4

Crowd Flow Segmentation

Identifying crowd flow patterns and direction of crowd flow is a crucial step in monitoring crowded scenes. The mass-gathering events like rallies, marathon, parades, protests, festivals, and sports matches, etc. comprise movement of the crowd in a bounded arena such as city pavement, over-bridges, and narrow roadways. An important step in the analysis of these crowded areas is segmenting the dominant flow patterns and directions. A crowd scene may consist of multiple flow patterns and information of the number of flows, and their location is not known beforehand. This makes flow segmentation a difficult task. The challenge of the computer vision system is to capture precise motion information, which is based upon the representation of motion. The required motion representation should create long and reliable trajectories. Using these trajectories, the flow of the whole video can be described.

4.1 Introduction

Conventionally, optical flow is used to compute pixel-wise flow between two frames [Brox and Malik (2010); Lu et al. (2010)]. However, optical flow can not directly be used for motion representation due to some implicit shortcomings. For instance, optical flow is unable to extract long-range spatio-temporal motion patterns, which are useful in a large number of applications.

In high-density crowd scenes, extracting complete trajectories from the crowd motion is a challenging task. Because of the aforementioned shortcomings of optical flow, trajectory or tracklets based representations [Zhao et al. (2011); Shao et al.

(2014); Dehghan and Kalayeh (2015)] has been favored by many crowd scene analysis approaches. Even though it gives long and reliable trajectories for motion representation, it still has numerous weaknesses. For example, in [Dehghan and Kalayeh (2015)], trajectory extraction is imposed in the whole crowd video frame, which is not required considering that some portion of the video frame is occupied by buildings, walls, trees, etc. Thus, a pre-processing step is required to segment the foreground (crowd region) from the entire frame with the aim to minimize further tracking.

Moreover, the performance of most of the standard object detection-based tracking methods gets deteriorated in a dense crowd [Ali and Shah (2008)]. This is because the high-density crowd scenes introduce complex dynamics among individuals and involve very small size object with occlusion, which is impractical to track simultaneously.

In this chapter, we present the problem of crowd flow segmentation in high-density crowd videos. The presented approach solves this problem by following an unsupervised paradigm which develops an active contour-based trajectory clustering algorithm for the purpose of crowd flow pattern segmentation. Our approach as depicted in Figure 4.1 has two strong points: first, low computational complexity with updating procedure attains the need to develop a real-time applicable crowd analysis system. Second, the representation of the trajectory feature creates a robust descriptor for a trajectory. By adopting active contouring, we pull out the possibility of tracking only the crowd region rather track the entire frame. Moreover, it does not demand specific training or learning for flow segmentation.

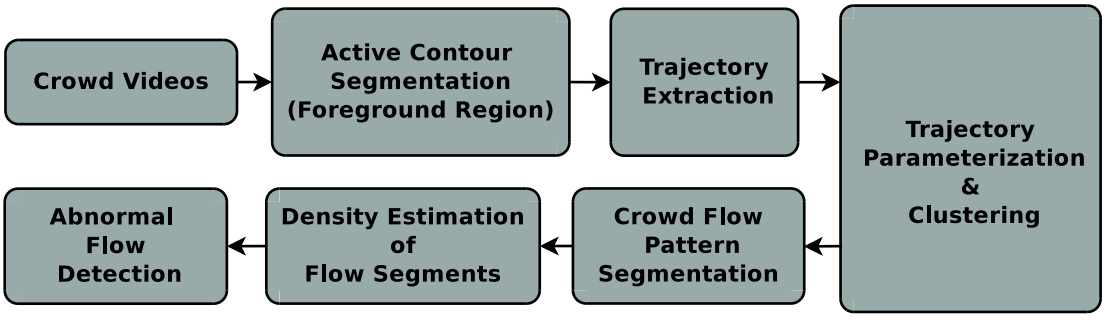

Figure 4.1: Schematic view of the presented method for crowd flow segmentation in high density crowd videos.

The remaining sections of this chapter are organized as Section 4.2 expounds our methodology to segment crowd flow patterns by trajectory clustering in the active contour region. Section 4.3 outlines the experimental results and comparative analysis of our approach with several state-of-the-art. Finally, the conclusions of

this chapter are given in Section 4.4.

4.2 Flow Segmentation Approach

Motion trajectories are an efficient manner to obtain and represent the complex temporal motion of a video scene. Therefore, we handle crowd flow segmentation through trajectory extraction and clustering task in the active contour region. Our approach consists of four phases:

1. Active contour segmentation: separate the crowd region from the entire frame.
2. Trajectory extraction: detect interest points in the block-level structure and track these blocks over the frames to extract the trajectories.
3. Trajectory clustering: cluster the extracted trajectories with the development of a clustering algorithm that utilizes shape, location, direction, and the neighborhood density of the trajectory patterns.
4. Crowd flow segmentation: mark individual pixels depending on their flow pattern to segment the final crowd flow.

Eventually, the crowd flow segments are analyzed to understand the flow. A detailed explanation of each phase is given in the consequent sections.

4.2.1 Active Contour Region Segmentation

Image segmentation plays a vital role in the detection, classification, scene understanding, and other visual analysis tasks. Active contour models (ACMs) have been broadly practiced for image segmentation. For acquiring precise results, this model utilizes prior knowledge related to image intensity distribution, boundary shape, and texture. ACMs can be classified as edge-based models [Caselles et al. (1993); Vasilevskiy and Siddiqi (2002)] or region-based models [Chan and Vese (2001)].

In edge-based methods, image gradient forces the active contours to shift toward the required boundaries of the given objects. These methods are prone to noise, and weak edges, having small gradient values, which may lead to edge leakage.

In region-based methods, image statistical information is utilized for forcing the active contours towards the borderline of the object. In several cases, this method is more beneficial than edge-based methods like in computer tomography (CT) and magnetic resonance (MR) images.

We adopt a region-based method that relies on the fact that in the high-density crowded scenes, the crowd region seems immanently homogeneous due to the same appearance from a distance. Since the similar region provides small changes in intensity while the inhomogeneous has a huge variation in intensity. Therefore, the difference between intensities offers a prominent feature to segment the foreground contour from the entire frame.

We follow CV model [Chan and Vese (2001)] for the contour segmentation with the aim to minimize the tracking domain by tracking only the segmented active contour region. Here, we focus on formulations where the video frame can be partitioned into a foreground region C_{in} and background region C_{out}, and the intensity distributions in different partitions are assumed to be independent. The contour is generated recursively in an unsupervised manner to minimize the energy function, and the minimum is obtained when evolving contour reaches the boundary between foreground and background regions.

The energy function of the CV model for a scalar-valued image is

$$\begin{aligned} \mu \cdot Length(C) + \vartheta \cdot Area(C_{in})) + \lambda^{+} \int_{C_{in}} \left(I\left(x\right) - C^{+}\left(C\right)\right)^{2} \\ +\lambda^{-} \int_{C_{out}} \left(I\left(x\right) - C^{-}\left(C\right)\right)^{2} \end{aligned} \tag{4.1}$$

where C is a contour, $I\left(x\right) \in R$ represents image intensity at x location of a pixel in spatial domain Ω. $\mu \geq 0$ denotes as a regularization parameter which controls the smoothness of the contour. The foreground and background regions are represented by C_{in} and C_{out} respectively. $\vartheta \geq 0$ is another regularization parameter, which penalizes a large area of the foreground. $C^{+}\left(C\right) = average\left(I\left(x\right)|x \in C_{in}\right)$ and $C^{-}\left(C\right) = average\left(I\left(x\right)|x \in C_{out}\right)$ represent the average intensities of the foreground and the background respectively. The parameter $\lambda^{+}, \lambda^{-} \geq 0$ controls the influence of the two image energy term $\lambda^{+} \int_{C_{in}} \left(I\left(x\right) - C^{+}\left(C\right)\right)^{2}$ and $\lambda^{-} \int_{C_{out}} \left(I\left(x\right) - C^{-}\left(C\right)\right)^{2}$ respectively at inside and outside of the contour.

In the variational level set formulation of Equation 4.1, the contour C is expressed as the zero level set of an auxiliary function $\phi : \Omega \rightarrow R$:

$$C := \{x \in \Omega : \phi(x) = 0\} \tag{4.2}$$

From Equation 4.2, foreground (C_{in}) and background (C_{out}) of contour C can be separately represented as in Equation 4.3

$$C_{in} := \{x \in \Omega : \phi(x) > 0\}, \quad C_{out} := \{x \in \Omega : \phi(x) < 0\} \tag{4.3}$$

The segmented foreground region, as shown in Figure 4.2) of the frame, is further processed for spatio-temporal interest points detection. The detected interest points are tracked over the frames. Here, according to our method, we refresh the temporal window after a certain number of r frames, then the existing contour region is re-segmented after every temporal window of r frames.

Figure 4.2: (a) Sample video frame (b) segmented foreground region using active contour approach (c) the segmented region is further divided into blocks.

4.2.2 Trajectory Extraction

The videos, we are dealing with, are densely populated with the moving persons, where each individual occupies very few pixels. The conventional object detection and tracking algorithms are not practicable, as maintaining track of individual persons in such a high density is a cumbersome process. Thus, instead of focusing on individuals, we consider tracking a block-level structure in the segmented foreground region C_{in}. For this purpose, we divide the foreground into non-overlapping blocks of size $b \times b$. At the boundary points, some of the blocks contain both the foreground and the background region. We only consider those blocks as per Equation 4.4, which consist of more than $\varepsilon\%$ pixels of foreground region.

$$Block(i) = \begin{cases} select & \text{if } C_{in}(pixelcount(block(i))) \geq \varepsilon\% \\ discard & \text{otherwise} \end{cases} \tag{4.4}$$

For each block, we prefer interest points as the primary feature to examine the intricate crowd motion. The interest points exhibits distinct entities that can be efficiently tracked in crowded scene.

Therefore, we adopt the most widely used Harris corner detector [Harris and Stephens (1988)] to detect these interest points. Suppose a grayscale image is represented as $R^2 \rightarrow R$, the variation in image space at point (x, y) is shown by a convolved matrix M_s as shown in Equation 4.5

$$M_s = g\left(.,.,\sigma\right) \times \begin{pmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{pmatrix} \tag{4.5}$$

I_x and I_y represent derivatives of image intensity in x and y directions respectively. $g\left(.,.,\sigma\right)$ is the Gaussian kernel with variance σ

$$g\left(.,.,\sigma\right) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.6}$$

An interest point is detected by computing positive maxima of the corner function as in Equation 4.7

$$H = \det\left(M_s\right) - K \times \left(trace\left(M_s\right)\right)^2 \tag{4.7}$$

As we are concerned about spatio-temporal scale, thus we follow the work of [Laptev and Lindeberg (2003)] and generalize the above Harris corner detector to the space-time domain.

The convolved matrix of interest points in the spatio-temporal domain is defined as in Equation 4.8

$$M_s = g\left(.,.,.,\sigma\right) \times \begin{pmatrix} I_x^2 & I_x I_y & I_x I_t \\ I_y I_x & I_y^2 & I_y I_t \\ I_t I_x & I_t I_y & I_t^2 \end{pmatrix} \tag{4.8}$$

where I_x and I_y represent derivatives of image intensity in x and y directions respectively and I_t denotes the derivative in temporal domain or consecutive frames

in time. The spatio-temporal interest points are chosen by their power which is defined by the corner function as in Equation 4.9

$$H = \det(M_s) - K \times (trace(M_s))^3 \tag{4.9}$$

where K represents the strength of spatio-temporal interest points. The more the variation in space and time, the interest point has the larger value of k as expressed in Equation 4.10

$$k = \frac{\lambda_1 \lambda_2 \lambda_3}{(\lambda_1 + \lambda_2 + \lambda_3)^3} \tag{4.10}$$

where λ_1, λ_2 and λ_3 are the eigenvalues of M_s given in Equation 4.8. Our concern is only for positive k and in this case, all the eigenvalues must be greater than 0 for space-time interest points.

After the interest points detection, for each block, a centroid of key-points is calculated. With the help of Kanade-Lucas-Tomasi (KLT) tracking algorithm [Tomasi and Kanade (1991)], the centroid of each block is then tracked over the successive frames. The tracker captures the trajectories for each moving block. The new objects simultaneously enter into crowd videos over time, some of them disappear after less than 20 frames and then reappear, which results in very short trajectories (tracklets).

To take into account the newly appearing persons for tracking, we refresh the frames after every r numbers of frames. Therefore, the existing foreground contour region and interest points are recomputed. When the blocks exit the frame, they are discarded from the tracker, and newly constructed blocks are considered for further tracking. We keep only the trajectories having a length greater than a certain threshold (LT).

The step by step process of trajectory extraction in the active contour region is summarized in Algorithm 1.

The visual representation of the extracted trajectories of some sample videos from datasets is depicted in Figure 4.4(b). After tracklets extraction in a temporal window of r frames in video sequences, we parameterize the trajectories as discussed in following section.

Algorithm 1 Tracklets extraction in active contour region

Require: Video Frames $(I_1, I_2, ..., I_n)$ Frame Refresh Rate (r)= 40, Length Threshold (LT)= 30;
Ensure: Extracted Tracklets

for i= 1 to n **do**
  if $(mod\,(i, r) == 0 | i == 1)$ **then**
    C_{in} = foreground segmentation as per Equation 4.3
    divide segmented foreground region C_{in}into blocks $b_1, b_2...b_m$ of size $(b \times b)$
    for j= 1 to m **do**
      if $C_{in}(pixelcount(block_j)) \geq \varepsilon\%$ **then**
        I_{pts}= interest points detection using Harris corner
        Centroid(I_{pts})= Centroid computation of set of I_{pts}
      else
        discard the block
      end if
    end for
  else
    Tracklets($k = 1, 2, 3...t$) = track all block's Centroid(I_{pts}) by KLT Tracker
  end if
end for
for all tracklets **do**
  L_k= Length computation
  if $L_k \geq$ LT **then**
    store or plot tracklets
  else
    discard the tracklets
  end if
end for

4.2.2.1 Representation of Trajectory Features

Representation of trajectory is one of the key factors in trajectory based motion segmentation. We represent the trajectory by taking into account the four features, which include shape, location, flow direction, and density features. The concatenation of these features creates a descriptor for a trajectory.

Suppose, a trajectory T is represented by $T = f\left\{(x_s, y_s), (x_{s+1}, y_{s+1}), ..., (x_e, y_e)\right\}$ where (x_j, y_j) denotes tracklet's coordinates at j^{th} frame having s as starting and e as ending frame. The shape, location, flow direction and neighborhood density features of the trajectory are defined as:

1. **Shape:** We capture the spatio-temporal shape of trajectory by expressing them as a third order polynomial functions of time. We separately represent

its shape in x and y dimensions as in Equation 4.11

$$\begin{aligned} x(t) &= p_0 + p_1 t + p_2 t^2 + p_3 t^3 \\ y(t) &= q_0 + q_1 t + q_2 t^2 + q_3 t^3 \end{aligned} \tag{4.11}$$

where $t \in (1, l)$ and $l = e - s$. The shape feature f_s denotes trajectory shape represented as $f_s = [p_0, ...p_3, q_0, ...q_3]^T$.

2. **Location:** The location of a tracklet is defined by the mean of its x and y coordinates represented by location feature vector, $f_{loc} = [\bar{x}, \bar{y}]^T$.

3. **Flow direction:** This feature decides the flow direction of trajectories in a video. Here, we determine 8 flow directions ranging from 0 to 360 degree. Each trajectory is mapped to a scalar variable f_m, which can take the value of 1 to 8 depending on its direction.

4. **Density:** We extract a multi-scale density feature from each trajectory. A trajectory obtains the information from its neighborhood trajectories, which is useful in segregating spatially overlapping trajectories. For a trajectory T_j, its density in a neighborhood of radius ϵ is computed as shown in Equation 4.12

$$\eta_{j,\epsilon} = \left| \left\{ T^i | \forall i \neq j, d\left(f^j, f^i\right) < \epsilon \right\} \right| \tag{4.12}$$

where $f^j = \left[f_s^j, f_l^j, f_m^j\right]$ represents concatenation of shape, location and direction features.

We compute a multi-scale density feature for the three values of epsilon to improve the robustness. Our multi-scale density feature f_d is given by $f_d = [\eta_{j,\epsilon_1}, \eta_{j,\epsilon_2}, \eta_{j,\epsilon_3}]^T$. The final feature vector F^j of a trajectory T_j is build by concatenating all the corresponding feature vectors extracted from a trajectory, which is represented by Equation 4.13.

$$F^j = \begin{bmatrix} f_s^j \\ f_l^j \\ f_m^j \\ f_d^j \end{bmatrix} \tag{4.13}$$

Likewise, all the trajectories form their feature vector as described in Equation 4.13, and next, we cluster the trajectories as explained in the following section.

4.2.3 Trajectory Clustering

Trajectory clustering provides benefits for many vision tasks, such as motion segmentation, object detection, action recognition, and scene modeling. Trajectories often lie in low dimensional feature space before clustering. The purpose of clustering trajectories in our approach is to segment flow patterns in a crowded scene.

In this section, we perform clustering using the similarity measures of trajectory feature vector F^j. We would like trajectories to be clustered together, which are belonging to the same flow segments. The cluster formation is a two-step process: firstly, all input trajectories are partitioned into a set of primitive clusters $C = \cup C_n$ where $n = 1 \cdots K$, then the primitive clusters belonging to same flow segment are grouped into N clusters.

4.2.3.1 Partitioning Trajectories into K-Primitive Clusters

We use the standard K-mean clustering algorithm [Likas et al. (2003)] to create a set of primitive trajectory clusters. The K-mean requires the value of the K parameter apriori. Usually, the value of K is larger than the predicted flow segments in a video. However, determining the value of K is not critical because the primitive clusters obtained from the K-mean algorithm will eventually merge into clustered trajectories based on minimum pairwise distance.

Firstly, the K trajectories are randomly chosen as initial clusters center, and then the rest of the trajectories are assigned to their nearest cluster centroids as per the Euclidean distance metric. K-mean follows an iterative procedure until the cluster centers are stabilized. This process converges to a local minimum only, as different cluster centers can arise by different initializations. Finally, the primitive clusters are formulated and used as building blocks to construct final flow clusters.

4.2.3.2 Merging K-Primitive Cluster into N Clusters

Once the primitive clusters are obtained from K-mean clustering, we randomly choose a fraction of trajectories from each primitive cluster. These trajectories are considered as representatives of all trajectory flow patterns referred as models.

Each trajectory is then matched against all the models for acquiring a priority set for every trajectory. In the priority set, a binary matrix is constructed by

storing a similarity match (either 1 or 0) of each trajectory to all models based on a specified threshold.

The model similarity between trajectories is computed as given in Equation 4.14.

$$P_{ji} = \begin{cases} 1 & |T_j - M_i| \leq \Delta \\ 0 & otherwise \end{cases} \tag{4.14}$$

Then, we perform agglomerative clustering between priority set matrix, representing the pairwise distance between any two trajectories $d(T_i, T_j)$ using Jaccard distance measure as in Equation 4.15.

$$d(T_i, T_j) = \frac{|M_i \cup M_j| - |M_i \cap M_j|}{M_i \cup M_j} \tag{4.15}$$

where M_i and M_j denote two sets of models corresponding to trajectory T_i and T_j.

Taking into account each trajectory as a cluster, we continuously merge the primitive clusters which have the minimum Jaccard distance at every iteration. After merging the minimum distance clusters at every iteration, we update the priority matrix and compute Jaccard distance for the remaining clusters. Repeat the process until the least distance between any two clusters is greater than δ.

This way, we cluster the trajectories belonging to the same flow patterns. The algorithm is summarized in Algorithm 2. The clustered trajectories of example video sequences are demonstrated in Figure 4.3.

From Figure 4.3, we observe that the mentioned trajectory clustering method over-clusters the given set of trajectories. To address this issue, we perform a second round of clustering over the previously clustered trajectories, as discussed in the following section.

4.2.4 Crowd Flow Segmentation

From the previous section, it can be observed that some of the trajectory clusters which comprise one flow segment are structurally different. This kind of situation occurs when people with the same flow direction comes into the scene at a different

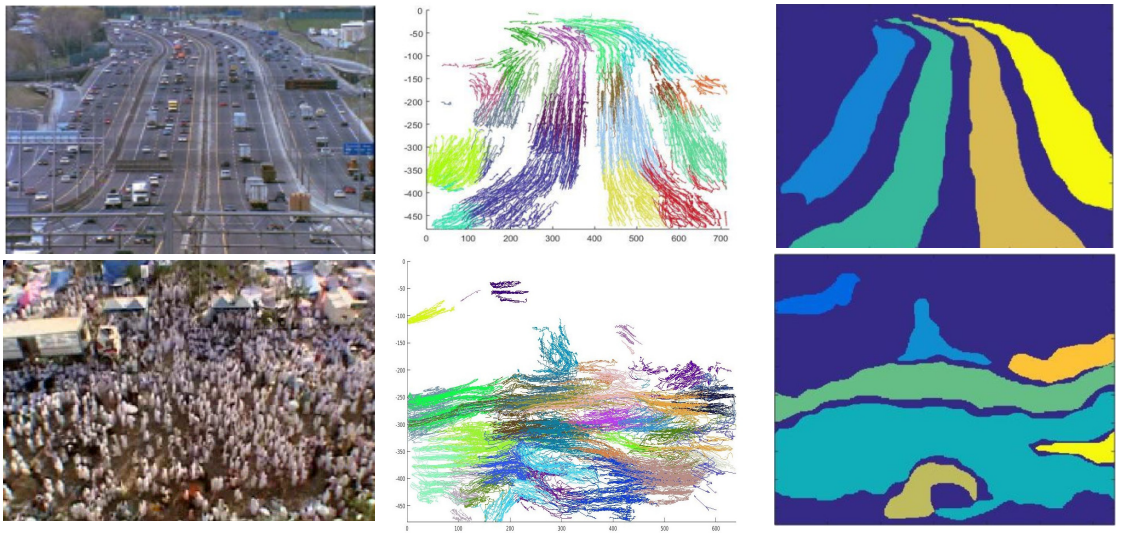

Figure 4.3: Results of trajectory clustering algorithm:(a)sample video frames (b) over clustered trajectories (c) and corresponding ground truths of crowd flow segments

Algorithm 2 Trajectory clustering algorithm.

Require: All Trajectory Features($F_{1,2,3,4..t}$)
Ensure: Trajectory Cluster $C_1, C_2, ...C_N$
  for All Trajectory features $F_{all(t)}$ **do**
    K-clusters (centroid) = Kmean(F_t, K)
    Models = Select-models(K-cluster, $F_{all(t)}$)
    P = Priority-Sets(Models, $F_{all(t)}$) using equation 6
    PW-dist = Pairwise-distance(P, P, Jaccard)
    Clusters(trajectories) = Clubbing Trajectories(PW-dist, P)
  end for
  for all i= 1:trajectories **do**
    $C_i = [i]$
  end for
  while min(PW-dist) < 1 **do**
    [i, j] $\leftarrow$ Search-position(min(PW-dist))
    $C_i = [C_i, C_j]$ Merge Clusters using agglomerative clustering approach
    Discard C_j
    Revise $P(i)$
    Discard $P(j)$
    Revise PW-dist
  end while
  RETURN $\boldsymbol{C_1, C_2, C_3, ..., C_N}$

time (i.e., they have trajectories with different start and endpoints). We customize the density-based clustering algorithm [Ester et al. (1996)] to resolve this problem.

The primitive clusters are taken as input to our flow segmentation algorithm. To start with, we randomly select cluster C_i as initial cluster and define its neighbor-

hood N_i which consists of $C_j(j \neq i)$ clusters. The cluster C_i and its neighborhood N_i should satisfy the following three paradigms.

1. The spatial overlapping between cluster C_i and C_j should be greater than α.
2. The variation in flow direction between cluster C_i and C_j should not go beyond the value of β.
3. The mean locations of a cluster C_i and C_j should not be far apart than the value γ.

Based on the aforementioned paradigm, the cluster C_i's neighborhood is computed and merged with all clusters of neighborhood N_i with cluster C_i. Then, we expand every neighborhood cluster one at a time and mark it visited.

Repeat the process for every newly included cluster until the neighborhood is processed. The unprocessed clusters are restored and follow the same procedure. Algorithm 3 explains the step by step procedure of final crowd flow segmentation.

The flow segmentation results are presented in the pixel domain, as shown in Figure 4.4(e). The pixels of trajectories which belong to one flow are combined as one region. Further, the sliding neighborhood operations are performed to smooth the boundaries of flow segments, as illustrated in Figure 4.4(f).

4.3 Experimental Results and Analysis

In Section 4.2, we presented the crowd flow segmentation method that has been formulated with the primary objective of precisely segmenting the crowd flow in the high-density crowd. We claim that by segmenting foreground contour and by clustering trajectories belonging to the same flow accomplish a comprehensive, optimal result achieving higher flow segmentation efficiency. To establish this claim analytically, we apply our approach on the distinct databases and evaluate those experimental outcomes.

In the following subsections, we outline parameter details and different performance measures used to evaluate the performance of the presented approach. Next, we discuss the qualitative and quantitative results of our method along with comparative analysis with other state-of-the-art on publicly available datasets.

Algorithm 3 Crowd flow segmentation

Require: Trajectory Cluster $C_1, C_2, ...C_N$
Ensure: Flow segments $S_1, S_2, ...S_f$
Initialization: unvisited cluster(C)= 0, f=0;
for all Trajectory cluster C_i **do**
Find neighborhood N_i
Increment in flow segment f
Call **ExpandCluster(**C_i, S_f, N_i**)**
Mark C_i as visited cluster
end for
function ExpandCluster(C_i, S_f, N_i**)**
Include C_i to segment S_f
for every C_j in N_i **do**
if C_j is unvisited **then**
Find neighborhood N_j
Merge neighborhoods (N_i, N_j)
Label C_j as visited
end if
if $C_j \notin S_f$ up to now **then**
$S_f \leftarrow S_f \cup C_j$
end if
end for
Return $S_1, S_2, ...S_f$

4.3.1 Parameter Details

A number of parameters are responsible for controlling the outcomes of the presented methodology. It is critical to understand and assign appropriate values to these parameters for generating better results. We decided the values of all mentioned parameters by empirical analysis. For this, we conducted our experiment over different values of thresholds on various video sequences of two datasets. We plotted an ROC curve and obtained an optimal value of the threshold. In this section, we discuss the parameters which are used in our work.

(a) After active contouring, as the obtained foreground region is divided into $b \times b$ size of blocks. Here to decide the size of a block, we conduct an experiment to evaluate how the block size can influence the trajectory extraction. When the block size is too small, it will result in pixel-based tracking of interest points, which is too noisy to track. The larger size of the block again leads to severe occlusion. In our experiments, we let the block size be 16×16 to obtain reliable trajectories.

(b) Next, the blocks which contain more than $\varepsilon = 25\%$ of foreground pixels are chosen to detect interest point, and track over the frames, i.e., the selection threshold is $\geq 25\%$. We decided it by empirical analysis. For this, we conducted our experiment over 20 different values of ε on 40 video sequence of two datasets. We plotted an ROC curve and obtained an optimal value of ε. Once the blocks are selected, they remain the same for $r = 40$ number of frames, and after that, the whole process in refreshed to take into account the newly appearing objects in the scene.

(c) For trajectory selection, we choose the length threshold $LT = 30$. Some objects move at very high speed and stay in camera for a very short period of time. These temporary objects do not contribute in obtaining long and reliable trajectories.

(d) In trajectory clustering, similarity match threshold $\Delta = 100$ and clustering termination threshold is $\delta = 1$. For flow segmentation, the value of α, β and γ is 10%, 20 and 150 respectively. Where α, β and γ represent threshold for spatial overlapping, difference in flow direction and mean locations between clusters respectively.

4.3.2 Performance Measures

To evaluate the performance of the presented method, we adopted three standard measures, which are the Jaccard similarity, F-score, and Mean Absolute Error (MAE). The Jaccard similarity measures the similarity between the two sets of data, as defined in Equation 4.16.

$$J(G, E) = \frac{|G \cap E|}{|G \cup E|} \tag{4.16}$$

Where G represents the ground truth of crowd flow segmentation for a video sequence and E be the experimental crowd flow segmentation obtained by the presented approach. The intersection counts the number of pixels that are shared between both labeled sets. The union counts the total number of segmented pixels occupied by any of the two labeled sets.

Next, we computed F-score as per Equation 4.17, which determines how well our approach performs. We consider the crowd flows correctly segmented if the segmented flow matches the marked ground truth region by more than 50%.

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.17}$$

We also computed Mean Absolute Error (MAE) for all sequences in our dataset, as defined in Equation 4.18. Where N represents the total number of flow segments in a video sequence, $z_i^{\cap}$ represents the number of segmented crowd flows, and z_i is the ground truth of crowd flow segments.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|z_i - z_i^{\cap}| \tag{4.18}$$

4.3.3 Datasets

Experiments are conducted on publicly available high-density crowd datasets which include UCF Web [Ali and Shah (2007)], Collective Motion [Zhou et al. (2013)] and some sequences of Violent Flows datasets [Hassner et al. (2012)].

The datasets contain varieties of graphics and real-life scenarios with a wide range of dynamics such as road traffic, marathon runner, railway stations, etc. For our research point of view, we considered only real life's crowd scenarios for evaluation. These scenes are suitable to be considered for crowd flow segmentation.

4.3.4 Qualitative and Quantitative Results

In this section, we show the experimental results and discussion of our work. The qualitative and quantitative comparison of presented work with other methods on each dataset is discussed in the following sections.

4.3.4.1 UCF Crowd Dataset

Our first experiment is performed on the UCF crowd dataset, which consists of 20 video sequences of graphics and real-life scenarios. We manually generated ground truths from all the video sequences and compared the presented approach with recent works of [Ali and Shah (2007); Biswas et al. (2014)] and [Kruthiventi and Babu (2015)]. Being dependent on videos, these methods are best suitable for comparative analysis. The qualitative comparison of our approach is limited

to only 4 videos because of the availability of the results of existing work [Ali and Shah (2007); Biswas et al. (2014)] and [Kruthiventi and Babu (2015)] for common videos is very small. The qualitative analysis of intermediate states involved in the presented approach is demonstrated in Figure 4.4.

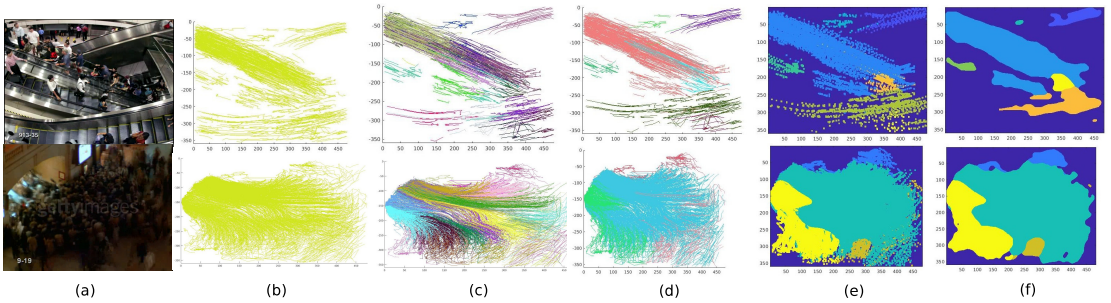

Figure 4.4: The main steps involve in our crowd flow pattern segmentation approach (a) example frames from crowd video sequences, (b) all extracted trajectories from videos, (c) primitive clustered trajectories obtained from K-mean clustering algorithm, (d) merged trajectory clusters resulting into crowd flow segments, (e) pixel wise crowd flow segments and (f) final crowd flows with smooth boundaries.

The comparative results for final flow segmentation of the presented approach are shown in Figure 4.5, demonstrating that we can handle complex crowd scenes much better than other existing methods.

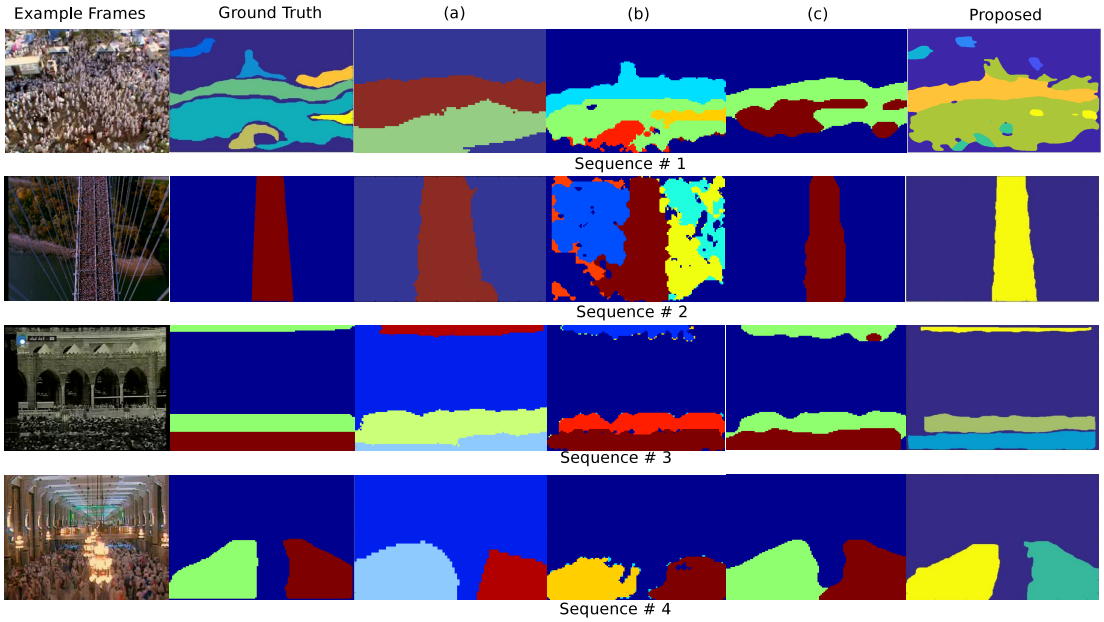

Figure 4.5: Comparative result analysis of our approach against existing work of (a) [Ali and Shah (2007)], (b) [Biswas et al. (2014)] and (c) [Kruthiventi and Babu (2015)] for crowd flow pattern segmentation on UCF Web dataset.

Most of the relevant methods [Loy et al. (2012)] do not share the implementation details. They only present qualitative results which are inadequate for present-

ing quantitative comparison among the state-of-the-art methods. The quantitative comparisons with other methods [Ali and Shah (2007); Biswas et al. (2014); Kruthiventi and Babu (2015)] by the values of Jaccard measure J are shown in Table 4.1 for some of the videos.

Our approach outperforms the benchmark method of [Ali and Shah (2007)] by the value of 0.84 of Jaccard similarity as we precisely segment the active contour region (foreground) of the video and discard the background region whereas [Ali and Shah (2007)] performs the flow segmentation at the global level by compromising the precision of the active contour region in the crowded sequence and obtained 0.57 of Jaccard similarity.

We also compared the presented approach with the work of [Biswas et al. (2014); Kruthiventi and Babu (2015)], and obtained better results as their approach does not handle intersecting flows due to mean statistics. [Kruthiventi and Babu (2015)] used conditional Random Field to model motion vectors, which are a computationally expensive method as compared to our approach and obtained 0.78 Jaccard similarity.

[Biswas et al. (2014)] proposed a super-pixel based algorithm in the compressed domain for flow segmentation and is outperformed by our algorithm with a large margin of 0.18 Jaccard dissimilarity for all video sequences.

It is observed that our approach can deal with intersecting crowd flow where the existing one [Ali and Shah (2007); Biswas et al. (2014); Kruthiventi and Babu (2015)] has failed. Therefore, our method outperforms the pixel domain and compressed domain approaches such as fluid dynamics by [Ali and Shah (2007)], super-pixel [Biswas et al. (2014)] and conditional random field of [Kruthiventi and Babu (2015)]. Our approach degrades only for high-speed videos where the objects stay for less than 5-10 number of frames in the video. For such cases, the extracted trajectories are very short, so they could not be included as flow.

Moreover, we present flow segmentation results of the above video sequences in terms of F-score, as shown in Table 4.2. We are unable to perform comparative analysis over this measure due to the unavailability of results of existing work over F-score measure . It can be noticed that our approach performs exceptionally well with only a few false flow segmentation, which is due to ambiguous or complex location motion.

Table 4.1: Quantitative analysis of our proposed approach (PA) with existing work on UCF Web dataset on Jaccard similarity measure.

Video Sequences	Jaccard Similarity Measure			
	Ali and Shah (2007)	Biswas et al. (2014)	Kruthiventi and Babu (2015)	PA
Sequence # 1	0.67	0.68	0.68	0.72
Sequence # 2	0.63	0.60	0.90	0.93
Sequence # 3	0.57	0.74	0.75	0.81
Sequence # 4	0.41	0.62	0.81	0.91
Average	0.57	0.66	0.78	0.84

Table 4.2: Quantitative performance of our flow segmentation (FS) approach on some video sequences of UCF Web dataset in terms of F-score measure.

Video Sequence	**# of GT**	**# of Correct FS**	**# of False FS**	**# of Miss FS**	**F-score**
Seq #1	7	6	2	1	.8
Seq #2	1	1	0	0	1
Seq #3	3	3	0	0	1
Seq #4	2	2	0	0	1

4.3.4.2 Collective Motion Database

We evaluate the presented approach on Collective Motion Database [Zhou et al. (2013)], which contains 413 video sequences of 62 crowded scenes. This database has various types of collective motions containing 100 frames per clips. We conducted our experiments on high-density real-world crowd sequences.

Our approach is compared with six state-of-the-art motion segmentation algorithms: The Lagrangian particle dynamics approach [Ali and Shah (2007)], the local-translation domain segmentation approach [Wu and San Wong (2012)], the coherent-filtering approach [Zhou et al. (2012a)], the collectiveness measuring-based approach [Zhou et al. (2013)], general motion segmentation method [Brox and Malik (2010)] and an anisotropic-diffusion-based image segmentation method [Wu et al. (2013)].

The qualitative comparison of our crowd flow segmentation approach with the aforementioned methods is demonstrated in Figure 4.6 for two video sequences. We include ground truths in the second column provided by [Lin et al. (2016)].

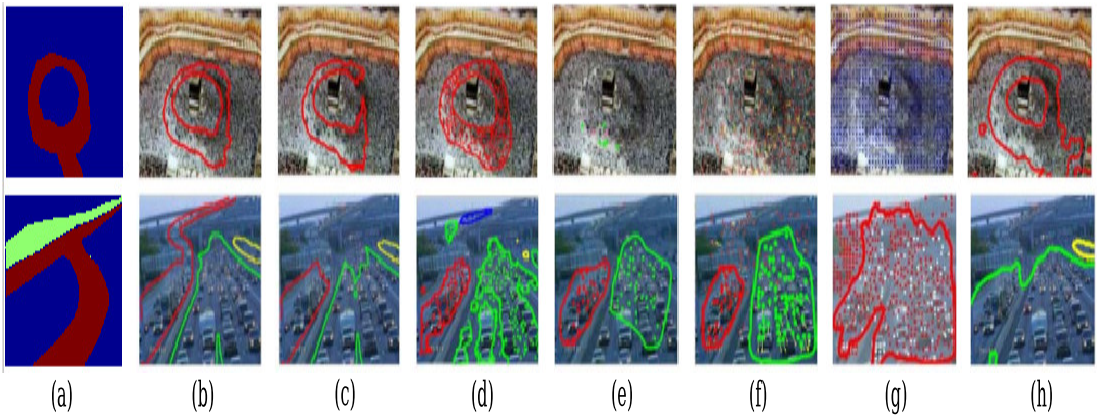

Figure 4.6: Crowd flow segmentation results: (a) results obtained from our approach, (b) manually mark ground truths [Lin et al. (2016)], (c) results of Lagrangian approach [Ali and Shah (2007)], (d) results of local-translation domain segmentation approach [Wu and San Wong (2012)], (e) results of coherent-filtering approach [Zhou et al. (2012a)],(f) results of collectiveness measuring-based approach [Zhou et al. (2013)], (g)results of general motion segmentation method [Brox and Malik (2010)] and (h) results of an anisotropic-diffusion-based method [Wu et al. (2013)]. (Best viewed in color)

It can be observed that the presented approach gets better segmentation results than the existing methods.

For instance, in the sequence 1, our approach effectively segments the circular flow while the Lagrangian particle dynamics approach by [Ali and Shah (2007)] can only segment the part of the circle. Moreover, the methods of [Zhou et al. (2012a)] and [Zhou et al. (2013)] fail to segment due to the extraction of only a few reliable key-points in the high density crowded scene.

Furthermore, for sequence 2, the methods in [Brox and Malik (2010)] and [Wu et al. (2013)] do not provide satisfactory results due to complexity of crowd scenes.

These methods fail to extract reliable particle flow or trajectories since boundaries of dense crowd flow are usually vague and unrecognizable.

The quantitative comparison of our method with other methods is reported in Table 4.3. The quantitative results of compared methods are given in terms of Mean Absolute Error (MAE), Jaccard similarity, and F-score.

Table 4.3 demonstrates that our approach achieves smaller values of MAE and higher values of Jaccard similarity and F-score than the other methods. This concludes that our approach can accurately segment the crowd flow in contrast to other approaches, which usually over-segment or under-segment the crowd flow regions.

Table 4.3: Quantitative comparison of our approach with other methods on Collective Motion Database in terms of Mean Absolute Error(MAE), Jaccard Similarity and F-score.

Methods	**MAE**	**Jaccard Similarity**	**F-score**
Ali and Shah (2007)	1.24	.75	.79
Wu and San Wong (2012)	0.93	.63	.70
Zhou et al. (2012a)	1.05	.43	.48
Zhou et al. (2013)	0.96	.47	.49
Brox and Malik (2010)	1.78	.37	.40
Wu et al. (2013)	0.84	.43	.44
Our approach	0.13	.82	.85

4.4 Chapter Summary

In this chapter, we presented an unsupervised approach for crowd flow segmentation in the active contour region. The active contouring method provides the separation between the crowd and the non-crowd region, which helps in the minimization of further tracking. We design a novel scheme for trajectory representation that utilizes shape, location, flow direction, and neighborhood density of trajectories. Further, we developed a trajectory clustering algorithm that employs a model-based grouping of trajectories. In contrast to existing methods, our approach can tackle complex crowd flows. In summary, our contributions to this chapter are

1. The approach minimizes the region to be tracked as compared to state-of-the-art techniques. We develop a foreground segmentation approach by using active contouring. Since the crowd is not present in the whole frame, some portion of the frame is occupied by buildings, walls, trees, etc. Therefore, the active contour approach segments the foreground crowd region from the entire frame to minimize further tracking.

2. We presented block-level tracking, especially suitable for high-density crowd videos due to the inadequacy of standard detection and tracking methods in the high-density videos.

3. The presented approach can handle intersecting crowd flows by designing a clustering algorithm that considers the shape, location, direction, and the

neighborhood density of trajectory patterns to cluster the trajectories. We exploit the essential temporal information depicted in trajectories for handling complex flow patterns.

4. We tested our approach on a set of benchmark datasets, and results are compared with several state-of-the-art methods. The performance is evaluated both quantitatively and qualitatively, and remarkable results are obtained.

Our approach yields very good results for high-density crowd scenes due to block-level tracking, which is a primary reason for decreasing the performance of standard tracking by individual detection approaches. As the presented approach does not require high computational resources, it can be used in segmenting flow patterns in real-time crowd video sequences. The primary objective of this work is to segment crowd flow while simultaneously considering the newly appearing object in the scene. It is accomplished by applying an updating procedure after a temporal window of a certain number of frames. Furthermore, the essential aspect of the presented method is taking advantage of foreground segmentation by an active contouring scheme.

The quantitative and qualitative results reported in Section 4.3.4 on standard datasets exhibit that the trajectory clustering method surpasses state-of-the-art methods used for flow segmentation in the high-density crowd. This algorithm can be utilized in other applications of the high-density crowd. These promising results motivate and enable us to analyze the flow segments and crowd scenarios in order to detect anomalous scene in crowd video, as detailed in Chapter 5.

Chapter 5

Anomalous Scene Detection

The purpose of anomalous crowd scene detection is early detection or prediction of anomalous events, which can result in life-threatening conditions for individuals. By early detection of anomalous scenes, potentially dangerous consequences can be reduced or prevented.

However, the analysis of crowd scene is a very strenuous task due to multiple occlusion between individuals, random variations in crowd density over time, low-resolution video sequences, complex background, and the inherent difficulty in accurately modeling the crowd behavior. The crowd behavior patterns which appear frequently are referred as normal pattern, and those appearing rarely are considered to as abnormal or anomalous patterns.

5.1 Introduction

The existing approaches to crowd behavior analysis can be broadly divided into object-based approaches and holistic approaches. The former approach considers crowd as a collection of individuals and goes through the challenges to recognize individuals in high-density crowd scenes due to a large number of targets, small resolutions, small target size, and severe occlusions, etc. Such circumstances lead to the loss of information of target objects in crowded scenes. To overcome this problem, some of the researchers, [Wang et al. (2009); Zhou et al. (2012b)] have adopted low-level features and probability models to analyze the dense crowd instead of focusing on tracking individuals. [Wang et al. (2009)] explored the hierarchical Bayesian model to segment motion into different activities by utilizing

visual features and atomic activities. [Zhou et al. (2012b)] classify different pedestrian behavior based on the mixture model of dynamic pedestrian, which learns the collective behavior patterns on the pedestrian. But, the mixture model considers affine transforms and faces the difficulty in representing complex shapes.

In the holistic approach, a crowd is considered as a global entity to judge the behaviors on the whole scene. These methods evaluate the dynamics of the whole crowd rather than focusing on the particular activity of each and every individual. Therefore, these techniques escape the exercise of individual detection and tracking people and explore crowd features to analyze the whole scene's behavior. This class usually incorporates an optical flow field-based approaches. [Benabbas et al. (2011)] developed a crowd model based on direction and magnitude and proposed a region-based segmentation algorithm which detects crowd event by learning different motion patterns.

Krausz and Bauckhage (2012) represented a global motion pattern by utilizing an optical flow histogram, which is used to detect stampede situations such as Love Parade stampede. [Rao et al. (2016)] developed an optical flow-based probabilistic framework using Riemannian manifolds to detect crowd activities.

In contrast to existing work that depends on motion cue merely between consecutive frames, more advanced approaches are developed that enhanced this cue to larger temporal frames by tracking some salient [Shao et al. (2014); Mousavi et al. (2015)] points or particles advections [Mehran et al. (2009, 2010); Mahadevan et al. (2010)]. These methods result in trajectories that capture more substantial temporal motion, which helps in to analyze crowd motion patterns.

In recent past, some researchers focused on particle advection approach [Mehran et al. (2009); Gu et al. (2014)]. In these approaches, a grid of particles is randomly distributed on the frame to represent crowd individuals and advected along with the optical flow. Here, particle resembles pixels, and the pixels are hard to differ with their neighborhood, which results in corrupted tracklets.

We address this issue in this chapter by considering the spatio-temporal interest points to be tracked. We filtered out the short length trajectories to not affect the persistent crowd flow direction. In this chapter, we describe an oriented tracklet's entropy-based approach for anomalous scene detection. In addition to entropy, we also compute temporal occupancy deviation that specifies the sudden huge changes in a crowd scene. We tested our approach on a set of benchmark datasets, and results are compared with state-of-the-art methods.

5.2 Methodology

In this section, an anomalous crowd scene detection approach is presented for high-density crowd events. Our approach performs oriented tracklets extraction and their statistical analysis in the active contour region to detect anomalous scenes. The approach consists of four main steps:

1. Active contour region segmentation: separate the foreground region from the entire video scene.
2. Tracklet generation: detection of spatio-temporal interest points blocks and track them over the frames.
3. Flow direction computation: compute tracklet directions and generate histogram of oriented tracklets.
4. Anomalous scene detection: entropy and temporal occupancy computation for oriented tracklets.

The anomalous crowd scene detection involves the deviation of a crowd from their normal behavior with respect to entropy and temporal occupancy. To classify a scene as anomalous, it should meet the following two conditions (i) the entropy of the scene is greater than the decision threshold, and (ii) temporal occupancy deviates from the specified bound limit. Further, an alert is issued to prevent crowd-related disaster.

The schematic diagram of our methodology is depicted in Figure 5.1 and, a detailed explanation of each step is given in the following sections.

5.2.1 Foreground Segmentation

We separate the foreground and background of the whole video frame by performing active contouring, as discussed in Chapter 4. We followed CV model [Chan and Vese (2001)] that defines two forces to partition the background and foreground, as shown in Equation 5.1.

$$\begin{aligned} & & F_1(C_{in}) + F_2(C_{out}) \\ \int_{C_{in}} |I(x) - c_1|^2 \, dx & + & \int_{C_{out}} |I(x) - c_2|^2 \, dx \end{aligned} \tag{5.1}$$

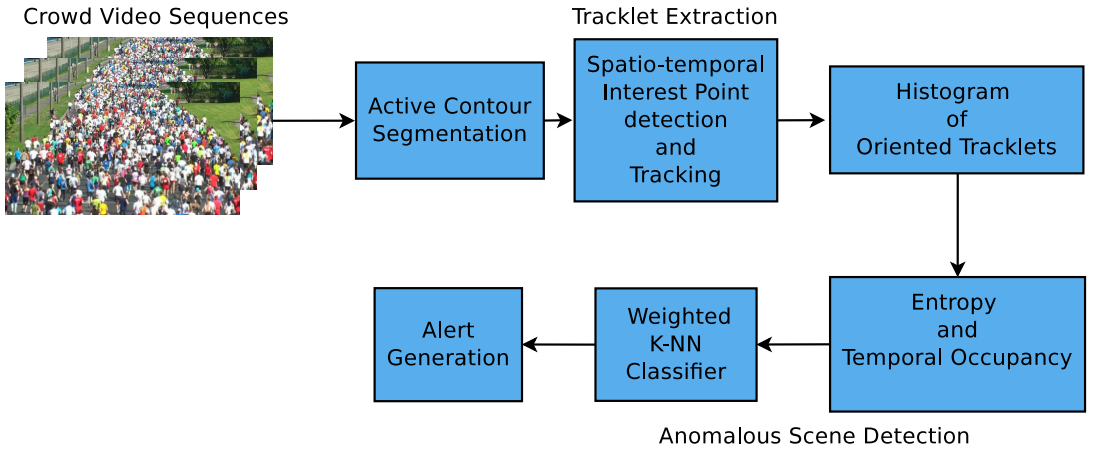

Figure 5.1: A schematic diagram of our approach for anomalous scene detection in high density crowd events.

The first force (F_1) is to shrink the contour and other one (F_2) is to enlarge the contour. As the contour touches the boundary between foreground and background regions, both the forces get stabilized. Here c_1 and c_2 denote the mean value of everything inside and outside of the contour respectively. The segmented foreground (C_{in}) and background (C_{out}) of contour C is separately represented as shown in Equation 5.2.

$$C_{in} := \{x \in \Omega : \phi(x) > 0\}, \quad C_{out} := \{x \in \Omega : \phi(x) < 0\} \tag{5.2}$$

The background region (C_{out}) is discarded since it does not contain the crowd, and further processing will be done in the foreground crowd region.

5.2.2 Tracklet Generation

We generate tracklets in space-time domain by utilizing spatio-temporal interest points, as discussed in Chapter 4. In our method, tracklet generation utilizes block-level structure in the segmented foreground region. Tracking a block-level structure is feasible in high-density crowd because maintaining track of individual persons in such a high-density crowd is a very cumbersome process. Tracking of interest points is also too noisy to maintain the tracks. Therefore, to generate the tracklets, the foreground C_{in} is partitioned into non-overlapping blocks of size $b \times b$. We discard those blocks, which consist of less than $\varepsilon\%$ pixels of foreground region. For each selected block, we detect spatio-temporal interest points by generalizing the work of [Laptev and Lindeberg (2003)] in spatio-temporal domain.

After the interest points detection, for each block, a centroid of key-points is calculated and then tracked over the successive frames by utilizing [Tomasi and Kanade (1991)] tracking algorithm. The tracker captures the tracklets for each moving block. The complete procedure of tracklet generation is repeated for every r numbers of frames to take into account the newly appearing persons for tracking. We keep only the trajectories having a length greater than a certain threshold (LT).

5.2.3 Tracklet Flow Direction

After tracklets extraction in a temporal window of r frames in video sequences, we attain the direction features of the tracklets. A flow direction feature defines the orientation of tracklets. We consider the direction of tracklets into 8 angles with the range of $[1:8]$ angles. Then, a variable f_m is attached to each tracklet that can select the value of 1 to 8 corresponding to its tracklet direction. The direction of tracklets is computed by arctangent as shown in Equation 5.3. Here, Δx and Δy are gradient of a tracklet.

$$Angle = \lfloor (\arctan 2\,(\Delta x, \Delta y) \div \pi \times 4 + 4) \rfloor \tag{5.3}$$

The dominant flow direction in a video can be determined by the median direction of tracklets as a representative direction as in Equation 5.4. Further, we distribute the direction of tracklets into histogram bins.

$$f_m = median(Angle(T_{all})) \tag{5.4}$$

5.2.3.1 Histogram of Oriented Tracklets

We quantize the histogram into 8 uniform bins and all the tracklets are distributed into them according to their flow direction. The tracklets associated with angle 1 are distributed in the first bin, angle 2 in second bin and so on. Each bin at x-axis of histogram consists orientation of tracklets and y-axis defines the frequency of tracklets at their respective orientation. The value of the eight bins defines the eight flow direction of tracklets, as shown in Figure 5.2.

The orientation histogram is computed after every temporal window of r frames. Further, we compute the entropy of each bin as per Equation 5.5.

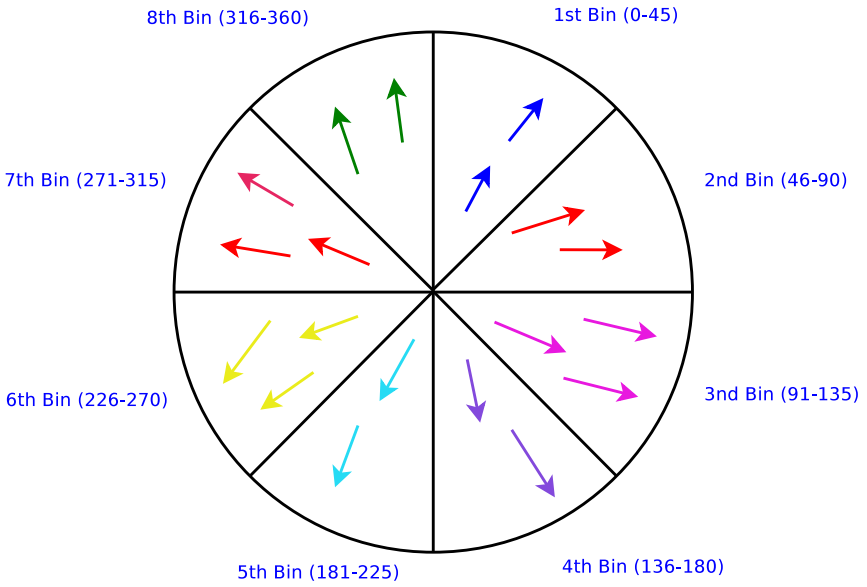

Figure 5.2: Flow direction of tracklets

$$\begin{aligned} E_{Bin_i} &= -P(Bin_i)\log_2 P(Bin_i) \\ P(Bin_i) &= \frac{Mag(Bin_i)}{\sum_{i=1}^{8} Mag(Bin_i)} \end{aligned} \tag{5.5}$$

Where $Mag(Bin_i)$ and $P(Bin_i)$ represent magnitude and probability of tracklets in i^{th} bin. Then summation of all is made to compute the final entropy of that temporal window as shown in Equation 5.6.

$$Entropy(final) = \sum_{i=1}^{8} E_{Bin_i} \tag{5.6}$$

5.2.4 Anomalous Scene Detection

Anomalous activities are usually categorized as an outlier detection problem. The definition of anomaly varies based upon the context. For instance, when an individual runs in the marathon is considered as non-anomalous, but the same scenario becomes anomalous for a road or mall.

In video sequences, the anomaly can be classified into two categories (i) local anomaly and (ii) global anomaly. In a local anomaly, small regions of the video act differently when compared to its nearest neighbors. For instance, the presence of a cart on a pedestrian path is a local anomaly, as shown in Figure 5.3(a).

In the case of global anomaly, the whole frame behaves differently as compared to its consecutive frame rather than neighbors within the frame. A frightening situation of the stampede is termed as a global anomaly, as depicted in Figure 5.3(b).

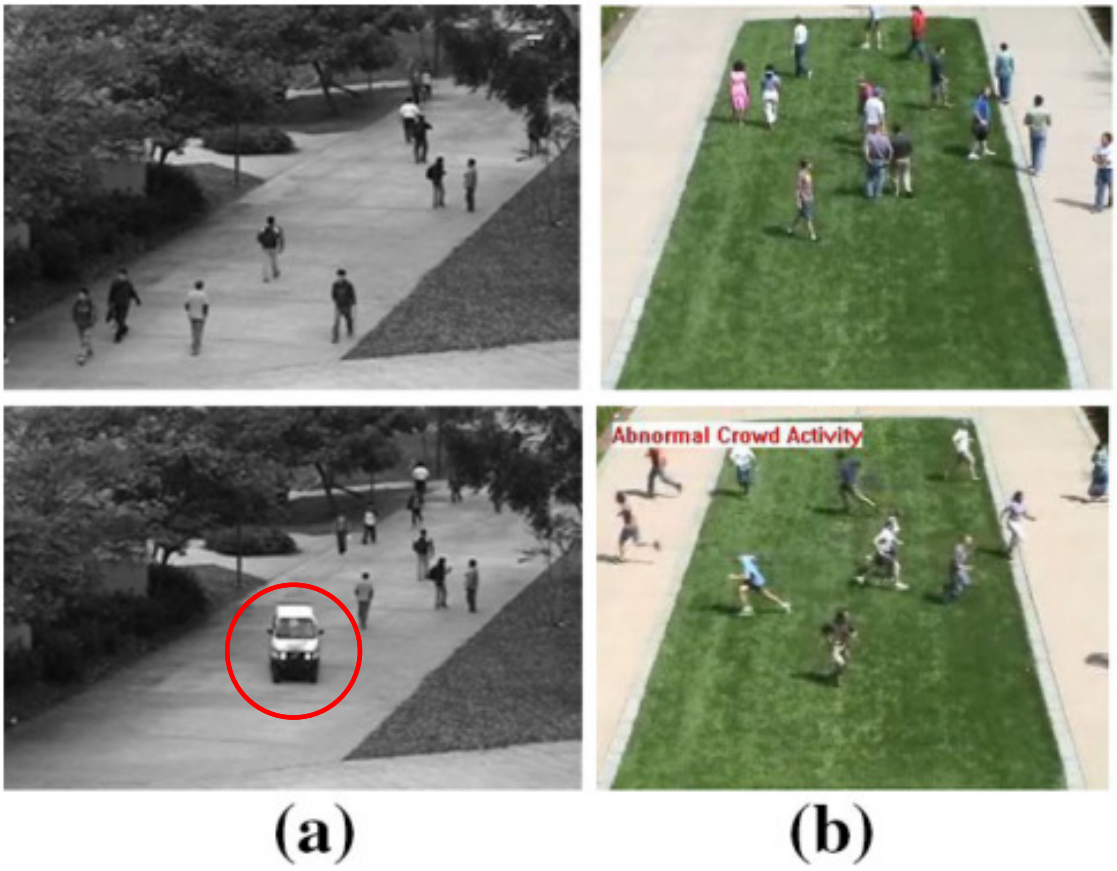

Figure 5.3: Illustration of global and local anomaly (a) Local anomaly: cart on the pedestrian path (b) Global anomaly: complete frame is anomalous as depicting a frighten situation.

We focus on global anomaly detection because in a crowd scene, what is happening is more important than who is doing it.

In this work, we consider the anomalous crowd scene detection by performing a statistical analysis on the oriented tracklets of the crowd. We adopt two statistical parameters as entropy and temporal occupancy for anomaly detection. Our idea is to model a normal crowd scene based on these parameters and any variation in the usual crowd behavior is detected as anomalous. The detailed description of these parameters is given in the following sections.

5.2.4.1 Entropy of Oriented Tracklets

The entropy reflects that infrequent events are more informative than frequent ones. Entropy provides a degree of uncertainty or randomness. Higher disorder or chaos leads to higher entropy. We present the entropy-based oriented tracklets approach to detect the anomalous scenes in the high-density crowd. Only the moving crowd can lead to anomalous activity; therefore, we considered only the moving blocks. The oriented tracklets are used to simulate the direction of the moving crowd.

Our approach intends to detect the anomalous behavior in crowd scenes according to the entropy measure of oriented tracklets. The temporal window will be considered as anomalous if its tracklets' entropy is higher than a certain threshold of τ. To compute the entropy of oriented tracklets, we distribute each tracklet to the histogram bins according to their corresponding orientation of flow direction. The probability distribution of tracklets in each bin is computed as in Equation 5.7

$$P_i(\theta) = \frac{T_i(\theta)}{T}, \quad i \in [1, 8] \tag{5.7}$$

Where $T_i(\theta)$ denotes the number of oriented tracklets in i^{th} bin at their corresponding angle and T represents the total number of tracklets in all bins. The entropy $E(t)$ of all oriented tracklets is computed as in Equation 5.8.

$$E(t) = \sum_{i=1}^{8} P_i(\theta) \log_2 \frac{1}{P_i(\theta)}, \quad i \in [1, 8] \tag{5.8}$$

where $P_i(\theta)$ represents the probability distribution of oriented tracklet at i^{th} bin. The entropy of extracted tracklets is measured for every temporal window of r frames, and if the entropy of the tracklets increases beyond a certain experimental threshold (τ), that means something of anomalous is happening.

5.2.4.2 Entropy based Scene Classification

Once entropy of a frame sequence is computed, we classify by predicting the class of a scene in real-time while minimizing zero-one loss (i.e., minimizing the number of false classification of an event and maximizing the concordance). We have used

weighed $k-NN$ classifier based on similarity measure for classification of an event in binary class as normal and anomalous.

Here entropy computed in Equation 5.8 is a predictor variable. Let $N_k(x)$ denotes the k nearest neighbors of current scene entropy (x) and, $\eta(x, x_l)$, $l = 1, ..., k$ denotes the similarity between current scene entropy (x) and a nearest neighbor sample (x_l).

A Gaussian kernel is used to compute the similarity measure between x and x_l as in Equation 5.9

$$\eta(x, x_l) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{d^2(x, x_l)}{2\sigma^2}\right] \tag{5.9}$$

where $d(x, x_l)$ is the Euclidean distance between the x and x_l. The similarity between x and x_l is normalized as expressed in Equation 5.10

$$z_l = \frac{\eta(x, x_l)}{\sum_{x_l \in N_k(x)} \eta(x, x_l)} \tag{5.10}$$

The classification of event x in class a is performed using the probability measure as:

$$P(y = a \mid x) = \sum_{x_l \in N_k(x)} \alpha_l z_l \tag{5.11}$$

where α_l is an indicator function represented as:

$$\alpha_l = \begin{cases} 1, & \text{if } y = a \\ 0, & \text{otherwise} \end{cases} \tag{5.12}$$

The weighted $k-NN$ classifier assigns event x to abnormal (i.e., class 1) if $P(y = 1 \mid x) \geq \tau$, otherwise a scene is classified as normal, where τ is used as the decision threshold.

5.2.4.3 Temporal Occupancy based Scene Classification

The temporal occupancy measures the area occupied by moving crowd blocks over time. To measure the occupancy of a scene, we estimate or count the number of tracklets in a scene. The occupancy is directly proportioning to the number of tracklets, i.e., more tracklets occupy more area. The occupancy of a scene can be determined by counting the persons existing in the scene. But, the conventional crowd detection and counting methods are cumbersome and less accurate in

densely crowded scenario due to high density, severe occlusion, and low resolution, etc.

We compute the tracklets of small windows corresponding to the moving crowd region and then count them at their respective orientation. The deviation in the number of tracklets in comparison to the previous temporal window is estimated as in Equation 5.13, which defines the percentage of area occupied (O) by tracklets during a time interval t_{th}.

$$TOD = \frac{O(t + t_{th}) - O(t)}{100}$$
$$TOD(t, t + t_{th}) = \begin{cases} true & TOD > \emptyset \\ false & \text{otherwise} \end{cases} \tag{5.13}$$

Where $\emptyset$ is a decision threshold for temporal occupancy. The massive deviation in the value of TOD indicates a sudden variation in the crowd scene. We can judge the abnormality of a scene concerning a massive deviation in temporal occupancy. This method is fast and shows effectiveness over template-based crowd detection or estimation methods.

The Algorithm 4 summarizes step by step process of anomalous scene detection by considering both techniques as mentioned above (i.e., entropy and temporal occupancy deviation).

5.2.5 Decision Threshold Selection

To estimate the decision threshold for anomalous scene classification, we adopted the receiver operating characteristic. The ROC estimates the optimal decision threshold for specified ground truths that meets the maximum concordance. Adjustment of decision threshold is utilized to enhance the performance of the sensitivity and specificity of a classifier under specified criterion.

We have analyzed three standard datasets for practical implementation and approximated various decision thresholds which fulfill the given sensitivity and specificity. Among experimented thresholds, the most appropriate threshold is selected corresponding to the highest area under the curve.

Algorithm 4 Anomalous Scene Detection Algorithm

Require: Tracklets $t_1, t_2, \cdots t_n$
Ensure: Anomalous or normal scene
  for all t_i= 1 to n **do**
    $t_i \rightarrow Bin_\theta \quad \forall \quad \theta \in [1, 8]$
  end for
  for all $Bin_\theta = 1$ to 8 **do**
    $P_t(Bin_\theta) = \frac{Mag(Bin_\theta)}{\sum_{i=1}^{8} Mag(Bin_\theta)}$
    $E_{Bin_\theta} = -P_t(Bin_\theta) \log_2 P_t(Bin_\theta)$
    E= $\sum_{i=1}^{8} E_{Bin_\theta}$
  end for
  for all Scene with frame interval equal to frame refresh rate (r). **do**
    Classify the scene based on two conditions given in Equation 5.11 and 5.13
    if $Entropy_f \geq \tau$ and
    $TOD(t, t + t_{th}) = true$ **then**
      Anomalous scene
      Generate alert
    else
      Normal scene
    end if
  end for

5.3 Experimental Results and Analysis

The goal of the presented approach is to analyze high-density crowd events that detect violent or anomalous behavior. Here, we outline parameter settings, performance measures, datasets and the comparative analysis of our method with the state-of-the-art on each dataset.

5.3.1 Parameter Settings

In this section, we discuss the parameters which are used in our experiments. In active contouring segmentation, once the foreground is segmented, the frame is divided into blocks of 16×16 size. The blocks which contain more than 25% of foreground pixels are chosen to detect interest points, and track over the frames, i.e., the selection threshold is $\geq 25\%$.

Once the blocks are selected, they remain the same for $r = 40$ number of frames.

For trajectory extraction: threshold for tracklet selection $t = 30$, frame refresh rate $r = 40$. For oriented trajectories, we consider the direction of tracklets in

the range of 0 to 360 degrees. The oriented tracklets are distributed into $n = 8$ number of bins with the interval of 45 degree.

In anomalous flow analysis, we have computed the entropy of oriented trajectories of various violent or anomalous and normal crowd scenes. With the diversity of scenarios, the optimal entropy for an anomalous scene is $E = 0.7$. For classification, the weighted K-NN classifier is used with $k = 5$. In anomaly detection, if the entropy increases beyond a certain decision τ value, then we generate an alert for anomalous crowd flow to prevent crowd-related disasters.

5.3.2 Performance Measures

For performance evaluation, we utilized both qualitative and quantitative measures. In qualitative analysis, we present the abnormal frames and their corresponding oriented tracklets and entropy in Figure 5.5. We discarded the short tracklets that disappear after less than 20 frames. The discrimination between normal and abnormal scene is made more clear after a set of frames and not a single frame. Thus, we computed the entropy of oriented tracklets after every r number of frames.

We also demonstrated some video frames which consist of both normal and abnormal scenes. For quantitative analysis, four performance measures are adopted as true positive rate (TPR), false-positive rate (FPR), receiver operating characteristic (ROC), and area under the curve (AUC). The true positive rate represents the rate of correct detection of an anomalous scene, and FPR is the rate of incorrect detection of an anomalous scene.

5.3.3 Datasets

For evaluation purpose, we used three publicly available crowd datasets which include UMN [UMN], UCF Web [Mehran et al. (2009)], Violent Crowd Flows [Hassner et al. (2012)] datasets and on some manually collected video sequences. The datasets contain varieties of scenarios with a wide range of dynamics such as road traffic, marathon runner, railway stations, etc. These videos are a collection of both synthetic and real-life scenes.

For our research point of view, we considered only real-life crowd scenarios for evaluation. Some sample frames of these datasets are depicted in Figure 5.4. Some

low-density crowd datasets are also considered for comparative study as most of the literature work is limited to the low-density crowd. Table 5.1 summarizes the statistics of all the datasets on which our approach is evaluated.

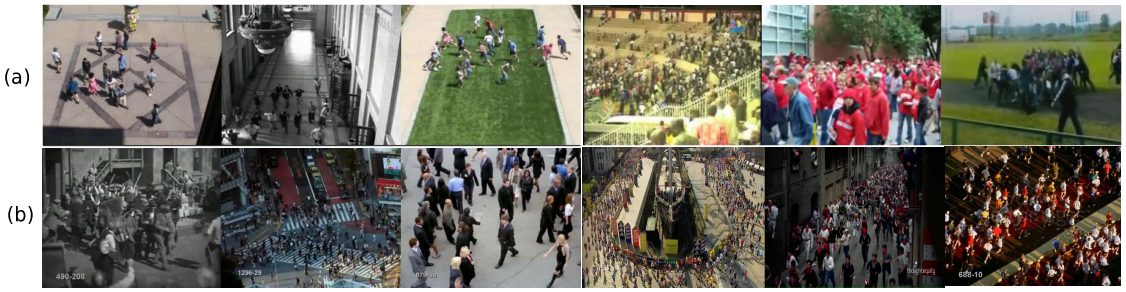

Figure 5.4: Sample frames of crowd datasets (a) first three coloumn represents UMN dataset while in next three Violent Flow Crowd, and (b) shows UCF Web and our collection of footage.

Table 5.1: Summarization of statistics of four crowd datasets namely UMN, UCF and Violent Crowd for anomaly detection, where NoF, FR, GT represents number of frames, frame rate and ground truth respectively. The last row represents our manually collected video sequences from the Internet.

Datasets	**NoF**	**Resolution**	**FR**	**GT**	**Scene/ Behavior**	**Density**
UMN Crowd [UMN]	7710	640 × 480	30	Yes	Real/Artificial	Low
Violent Flows Crowd [Hassner et al. (2012)]	221400	Varied	30	No	Real/Natural	High
UCF Web Crowd [Mehran et al. (2009)]	24480	640 × 480	24	No	Real/Natural	High
Collected Footage	24000	Varied	30	No	Real/Natural	High

5.3.4 Qualitative and Quantitative Evaluation

The qualitative result of intermediate steps of our approach is illustrated in Figure 5.5 on some example frames of normal and abnormal scenes. In Figure 5.6, an example video sequence, and their corresponding entropy graph are depicted for every temporal window of r frames. The sudden change in entropy value indicates that something anomalous has happened. Performance evaluation of our method on each dataset is discussed in the following sections.

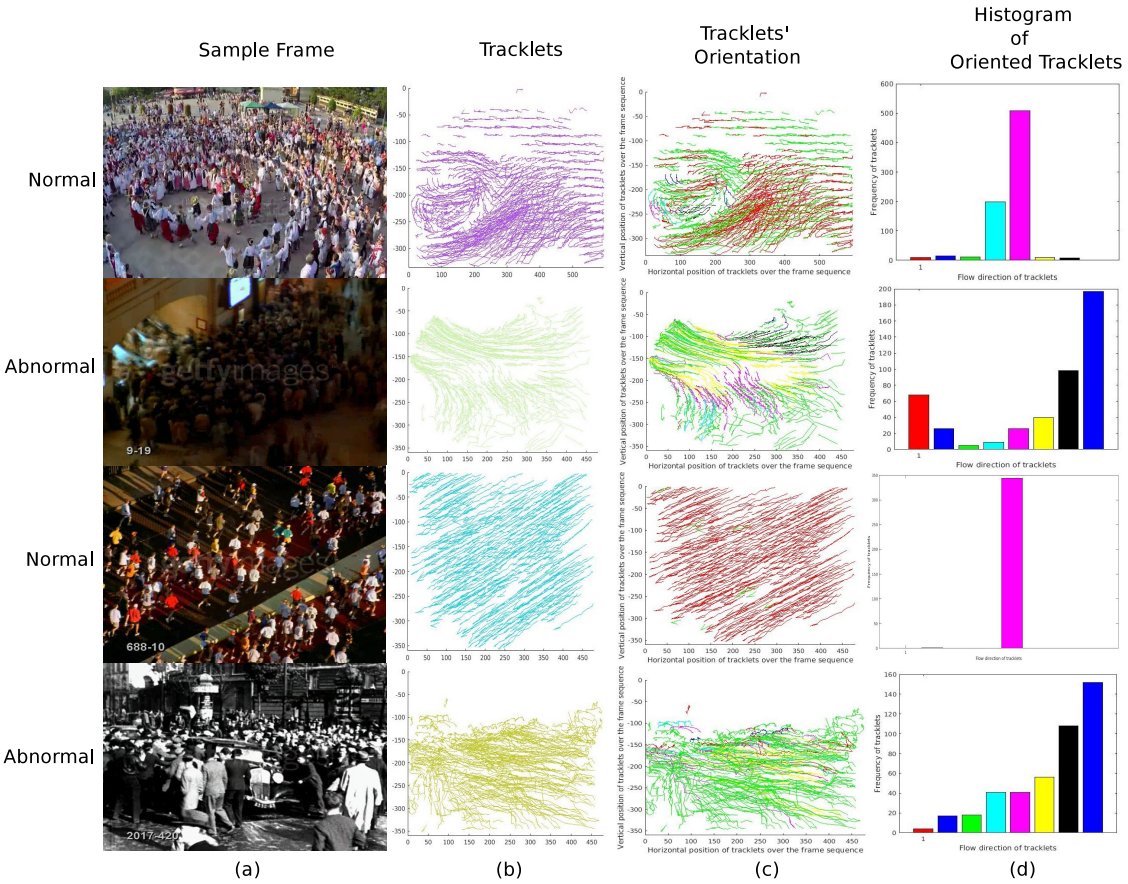

Figure 5.5: The qualitative representation of the intermediate outcomes of our approach on some sample normal and abnormal video sequences (a) shows example frame of normal and abnormal video sequences, and in (b) extracted tracklets are depicted (c) represents the flow direction of the tracklets and the corresponding histograms of the tracklets' orientation are depicted in (d). It is observed that the tracklet of abnormal sequences are spread-out in all or random directions while the direction of the tracklets of normal sequences produced dominant flow.

5.3.4.1 Evaluation on UMN Dataset

Our first study of experiments is on UMN dataset [UMN] which is publicly available on the website of the University of Minnesota[1] to figure out the performance of anomaly detection algorithms. The dataset resides 11 video clips of three different scenarios. Each video sequence commences with general or normal behaviors and ends up with a panic running crowd. It consists of 7740 number of frames with 320×240 resolution. Though the dataset is limited regarding crowd density and variations, but we included this dataset to consider the comparative study of existing literature.

The comparative results of our approach with the other state-of-the-art such as so-

[1] http://mha.cs.umn.edu/proj_events.shtml#crowd

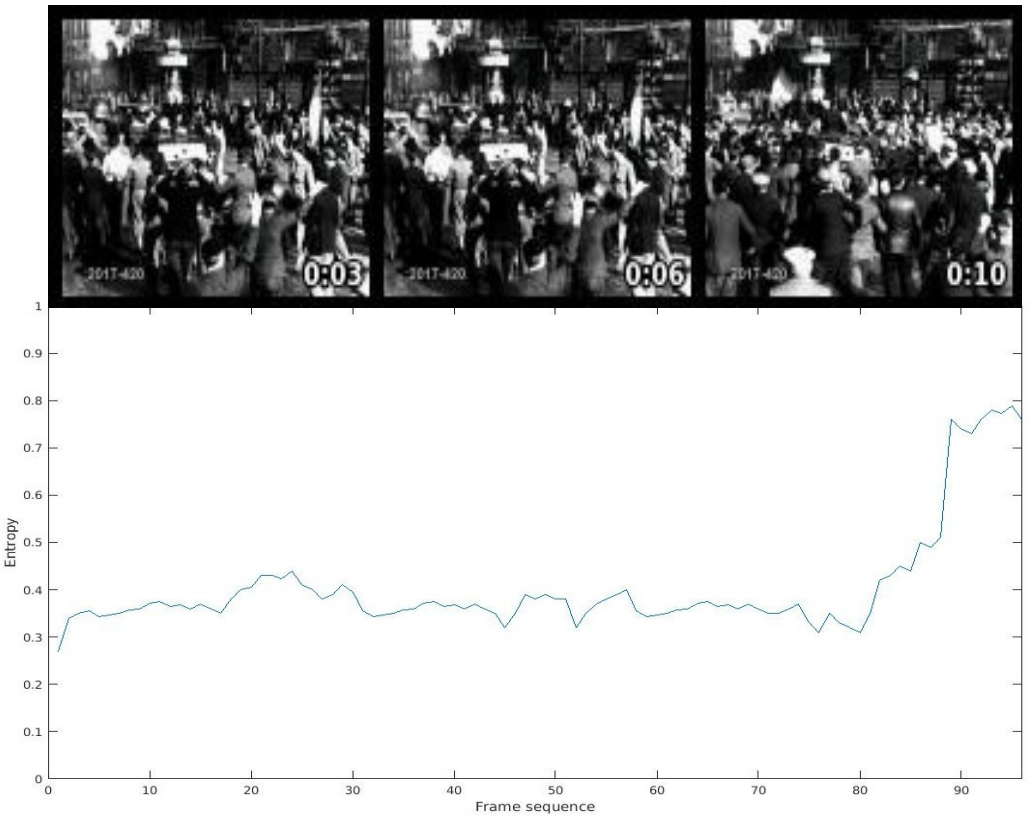

Figure 5.6: Some snapshots of a video sequence and their corresponding entropy graph of every temporal window of r frames. The sudden change in entropy values indicates that something anomalous has happened.

cial force model [Mehran et al. (2009)], optical flow [Mehran et al. (2009)], chaotic invariants [Mahadevan et al. (2010)], sparse reconstruction cost (SRC) [Cong et al. (2013)] and matrix approximation [Wang and Dong (2012)] methods are reported in Table 5.2.

All the quantitative results of compared approaches have been taken from [Cong et al. (2013)]. These are the best suitable methods for comparative analysis. [Mehran et al. (2009)] used the social force model to compute the interaction among the crowd. It detects anomalies based on a high deviation in the magnitude of the interaction force between particles. Their approach limits, when the video scene consists of high-density particles because the interaction force between neighbors will always remain high in a dense crowd. Their approach achieves good results only with the sparse crowd. They also used a pure optical flow-based method, which achieves 0.84 of AUC.

[Wang and Dong (2012)] compute motion matrix of video sequences, the motion-based methods degrade performance in the presence of noise. And, the noise and disturbances are obvious in natural crowd scenes.

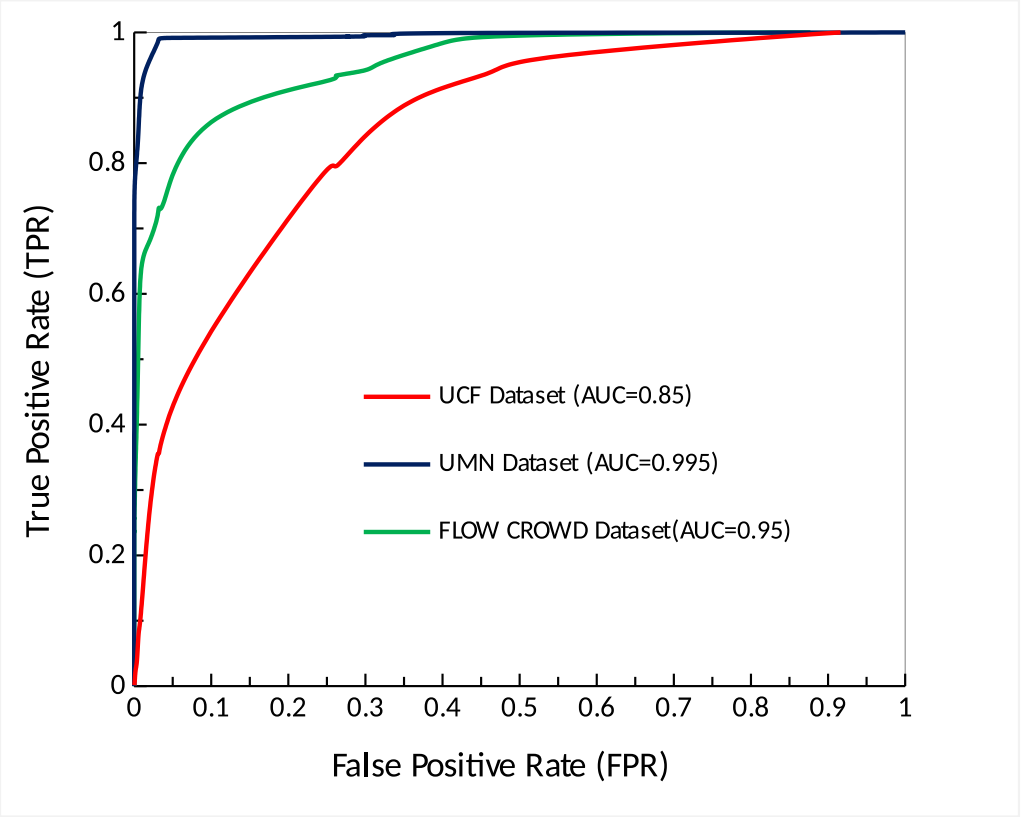

Figure 5.7: The ROC curves for each experimented datasets as UMN, UCF Web and Violence Crowd datasets.

[Mahadevan et al. (2010)] defines chaotic invariants to analyze a scene and [Cong et al. (2013)] utilizes the sparse reconstruction cost of motion field.

We observe that all methods, as mentioned above, detect the abnormality at a frame or pixel-level while our approach works at scene level as the distinction between normal and abnormal scene is more apparent after a certain interval. This data set produces significant changes in the motion pattern of the scene. Due to this noticeable change, our approach achieves nearly perfect results closest to the ground truth, i.e., we achieve 0.995 value of area under an ROC curve. The compared methods also hold good results but outperformed by our method as depicted in Table 5.2. The ROC curve of our approach on a different dataset is shown in Figure 5.7.

5.3.4.2 Evaluation on UCF Web Dataset

Our next evaluation is on the UCF web crowd dataset which has been used by [Mehran et al. (2009)]. This dataset has 20 videos of normal and abnormal crowd scenes with 640×480 resolution. The crowd is on the pedestrian walk or running in

Table 5.2: The quantitative comparative analysis of our approach with the state of the art methods for detection of an anomalous scene in the publicly available UMN Crowd dataset.

Methods	**AUC**	**Speed (FPS)**
Optical flow [Mehran et al. (2009)]	0.840	10
FCSB [Pennisi et al. (2016)]	0.950	30
Social force model [Mehran et al. (2009)]	0.960	10
Chaotic invariants LP [Mahadevan et al. (2010)]	0.990	15
Sparse reconstruction cost (SRC) [Cong et al. (2013)]	0.980	20
Matrix approximation [Wang and Dong (2012)]	0.980	10
Our approach	0.995	30

marathons is considered as normal behavior while the abnormal behavior consists of a stampede, protesters clashing, fighting and running in random directions. Out of 20 videos, 12 video sequences are normal and the remaining 8 are abnormal. For each footage, we manually marked the ground truth to abnormal frames.

For performance evaluation, we computed TPR, FPR, precision and also plotted the ROC curve as shown in Figure 5.7. We have compared the area under ROC curve with the social force model [Mehran et al. (2009)], optical flow [Mehran et al. (2009)] and GLCM [Lloyd et al. (2017)] methods and demonstrated in Table 5.3.

[Mehran et al. (2009)] analyze the abnormality by advecting a grid of particles into the scene and move them with the elemental optical flow field. Due to the similarity of pixels, the advected particles are hard to distinguish and results in corrupted flow. In the behavior of escape panic and sudden change, interaction forces exhibit contrary results. Also, their approach requires prior training, thus efficient for the real-time application.

[Lloyd et al. (2017)] proposed a GLCM texture-based anomaly detection method, which requires a strong structure that is scarcely available in real-world random crowd scenes. This dataset is not widely tested; therefore, our comparison is limited to only two approaches, as mentioned above, and our presented approach outperforms these two methods in terms of AUC. [Mehran et al. (2009)] holds 0.66 and 0.73 values as AUC for their optical flow and social force method respectively. The AUC value for [Lloyd et al. (2017)] is 0.82 which is much better than both the approaches proposed by [Mehran et al. (2009)]. Our approach achieves 0.85 value of AUC, is best amongst all the compared approaches.

Table 5.3: Quantitative analysis of our approach with the other state of the art methods on UCF Web Crowd Dataset for anomalous scene detection.

Methods	**AUC**	**Processing Speed (FPS)**
Optical Flow [Mehran et al. (2009)]	0.660	5
Social Force Model [Mehran et al. (2009)]	0.730	5
GLCM [Lloyd et al. (2017)]	0.820	10
Our approach	0.850	30

5.3.4.3 Evaluation on Violent Flows Crowd Dataset

Finally, we tested our approach on a highly complex Violent Flows Crowd dataset obtained from [Hassner et al. (2012)]. This dataset is recent and rarely adopted in the literature. The sole purpose of using this dataset is to evaluate the crowd abnormality and violence in scenes. The dataset consists of 123 samples of both violent and non-violent crowd scenes collected from the footage which are uploaded at various websites. The violent video consists of various sequences which give similar visual perceptions as in real-world crowd scenes. Therefore, this dataset is appropriate to evaluate the performance of our approach for anomalous scene detection in the high-density crowd.

We manually mark the frames which turn to violent from normal scenes by a majority vote of 20 persons. Different methods are compared by their AUC for anomalous scene detection. Specifically, we consider five methods [Hassner et al. (2012); Xu et al. (2014); Gracia et al. (2015); Gao et al. (2016)and [Lloyd et al. (2017)] for comparison with our approach.

In [Gracia et al. (2015)], fast fight actions are detected by blob processing, the methods merely focus on fight actions with low accuracy and are not suitable for general crowd violence flow.

The methods of [Gao et al. (2016)] and [Hassner et al. (2012)] also focus on violent scene detection by relying on the magnitudes of the optical-flow field. Both approaches require a trained model to classify the violent and non-violent scenes.

The author of [Xu et al. (2014)] captures distinctive local shape and motion patterns of activity to discriminate between violent and non-violent. The local shape of activities is occluded in high-density crowd scenes.

[Lloyd et al. (2017)] describe the visual texture of scenes by using the statistical properties of GLCM feature. The key requirement of their approach is to maintain

the strong structure of scenes which is not practical in real-time surveillance systems as the footage may be influenced by illumination effects, darkness, occlusion, etc.

In contrast to the above-aforementioned approaches, our approach does not require a pre-trained model and operates in real-time scenarios. Table 5.4 demonstrates the quantitative comparison of our approach with other state-of-the-art methods. And, it is observed that the presented approach outperforms the state-of-the-art with 0.95% value of the AUC as shown in Figure 5.7. The quantitative comparison of our method with other state-of-the-art methods on three different datasets is shown in Table 5.5.

Table 5.4: The quantitative comparative analysis of our approach with the state of the art methods for detection of the anomalous scene in the publicly available Violent Flows Crowd Dataset.

Methods	**AUC**	**Processing Speed (FPS)**
Fast Fight [Gracia et al. (2015)]	0.750	10
OViF [Gao et al. (2016)]	0.805	20
ViF [Hassner et al. (2012)]	0.850	30
MoSIFT [Xu et al. (2014)]	0.875	20
GLCM feature [Lloyd et al. (2017)]	0.940	10
Our approach	0.950	30

5.4 Chapter Summary

This chapter presents a novel method for anomalous scene detection in high-density crowd employing statistical analysis of oriented tracklets. We employ a tracklet extraction approach in active contour region. To extract tracklets, spatio-temporal interest points are detected and tracked over the frames. With the active contour method, only the foreground crowd region needs to be tracked that optimize the tracking phase. The orientation of tracklets is computed and quantized in histogram bins with respect to their flow direction.

Further, the entropy and temporal occupancy of oriented tracklets are measured at each temporal window. The huge deviation in both measures predicts the scene as anomalous. At the initial level, a weighted K-NN classifier is adopted that predicts the probability of class based on Gaussian similarity measures. If a scene is predicted as anomalous, an alert is issued to prevent the crowd-related disaster such as over-crowdedness and clogging.

Table 5.5: The quantitative analysis of our approach with the other state-of-the-art methods on three different datasets.

Methods	Area Under Curve (AUC)			Processing Time
	UMN	UCF	Violent	
Optical flow [(Mehran et al., 2009)]	0.660	0.840	–	10
FCSB [(Pennisi et al., 2016)]	0.950	–	–	30
Social force model [(Mehran et al., 2009)]	0.960	0.730	–	10
Chaotic invariants [(Mahadevan et al., 2010)]	0.990	–	–	15
SRC [(Cong et al., 2013)]	0.980	–	–	20
Matrix approximation [(Wang and Dong, 2012)]	0.980	–	–	10
GLCM [(Lloyd et al., 2017)]	–	0.821	0.940	10
Fast Fight [(Gracia et al., 2015)]	–	–	0.750	10
OViF [(Gao et al., 2016)]	–	–	0.805	20
ViF [(Hassner et al., 2012)]	–	–	0.850	30
MoSIFT [(Xu et al., 2014)]	–	–	0.875	20
Our approach	0.995	0.850	0.950	30

Experiments are conducted on three datasets: UMN, UCF Web, and Violent Flow. The UCF Web and Violent Flows datasets are complementary to the presented approach as they consist of dense crowd scenes of real-world applications such as the marathon, stadium, stampede, and political rallies, etc. The experimental results obtain promising results when compared to existing approaches. Our approach maintains low computational complexity and can be applied to real-world crowd video sequences to detect potentially anomalous scenes.

Chapter 6

Conclusions and Future Work

This thesis work addresses several important problems of crowd analysis in the high-density crowd with hundreds to thousands of people in an image or frame with the aim to ensure public security and safety. In this research work, we study various state-of-art methods used in the different tasks of crowd analysis and find their limitations in various challenging environments. Further, we presented techniques that aim to analyze the dense crowd in terms of density estimation, flow segmentation, and anomalous scene detection. The presented techniques can be implemented in any language that supports image processing, and it is not prescriptive of any particular platform or tool. In this chapter, we present the summary and conclusions drawn from our research study and the possible improvements as future directions to extend this work.

6.1 Summary of the Work

Chapter 2 provides a detailed literature survey of existing work for different tasks of crowd analysis. The chapter explores various challenges that exist in crowd analysis techniques. It also addresses the limitations of the widely used methods for each of these tasks. The chapter illustrated the current image and video-based crowd analysis techniques in multidisciplinary research such as intelligent surveillance systems, public space design, anomaly detection, etc.

In Chapter 3, we presented detection and regression-based approaches for density estimation in images of exceptionally dense crowds comprising an average of a thousand people per image. The complete description of the crowd detection

method using the skin color model, oriented gradient features, and the support vector machine has been presented. It is observed that face detection in the high-density crowd in the presence of severe occlusion, blur, and perspective effects remains an understudied area. We have shown some failure cases of face detection in the high-density crowd, which limits the applicability of detection based methods in such a challenging and high-density crowd images. This chapter further introduced a detailed description of regression-based approaches. These approaches fuse information from four different sources in terms of counts and confidences. This chapter also reported the comparative analysis of both approaches with state-of-the-art methods.

In chapter 4, we presented an unsupervised approach for crowd flow segmentation in the active contour region. The contouring helps in the minimization of further tracking. The crowd flow segmentation approach starts by proposing a trajectory clustering algorithm that uses a model-based grouping of trajectories that can handle complex motion patterns and intersecting crowd flows.

Chapter 5 concerns to detect anomalous scenes in high-density crowd videos. The process of implementation of the methodology describes the entropy and temporal occupancy as decision criteria of abnormality of the crowd video. We have formulated a method for abnormality detection by tracklet analysis. The orientation of tracklets is computed and quantized in histogram bins with respect to their flow direction. The entropy and temporal occupancy are computed for each bin of the histogram. Classification of an abnormal scene is performed via similarity measure with the normal behavior of the crowd scene. The normal crowd scenes are modeled on statistical parameters, and any inconsistency in the normal crowd behavior is identified as anomalous. The presented techniques can be applied to any high-density crowd video and do not limit to specific video class.

6.2 Conclusions

In this work, we present an efficient system to analyze crowd scenes. We have explored different tasks involved in crowd analysis and optimization is performed over each task.
The major contributions of our research work can be summarized as follows:

1. We have found that the task of crowd density estimation remains challenging when there are a high density of the crowd, severe occlusion, and perspec-

tive effects, etc. We formulate a texture-based multi-source approach which leverages the clues obtained from different sources, to figure out the density of people existent in exceptionally dense, crowded regions with repeated texture. Our experimental results validate the adequacy and efficiency of the intended methodology by achieving remarkable performance in terms of MAE and MSE when compared with existing methods of crowd density estimation.

Following implementations obtain these improvements:

(a) We grasp the multiple information gathered from various sources such as Fourier, Wavelet, head and SIFT in terms of confidence scores and different statistical parameters instead of using a single feature.

(b) All these techniques are sequentially applied in local neighborhoods (patch-level) at multiple scales for avoiding the issues of foreshortening and local geometric distortion arising from irregularity in the observed textures rising from dense crowd images.

(c) Smoothness and consistency are maintained among neighboring patches by employing Markov Random Field.

2. We have found that feasible solutions are available for addressing the robust tracking of the single object. However, simultaneous tracking of individuals in high-density crowd scene stays as one of the most challenging tasks in crowd analysis. The accuracy of standard tracking methods reduces as the density of crowd and occlusion between them increases beyond hundreds of people. To achieve high performance in crowd tracking to segment crowd flow patterns, we have applied the following enhancement:

(a) We segment foreground (crowd region) with the aim to minimize tracking that takes segmented foreground regions as input to select features for further tracking.

(b) We track mid-level structure (block-level tracking), which is especially suited for high-density crowd videos because individual detection and tracking do not work for high-density crowd videos.

(c) The performance of crowd flow segmentation mainly depends upon the accurate selection and representation of the trajectory feature of the moving crowd. We aim at the robust selection and representation of trajectory features. We exploit the rich temporal information contained in trajectories for handling complex flow patterns.

(d) We also continuously take into account the newly appearing moving objects in a video by revising our existing set of tracker constituents after processing a temporal window of frames.

(e) We perform trajectory clustering considering the shape, location, direction, and the neighborhood density of trajectory patterns to cluster the trajectories for flow segmentation.

We test our approach on a set of benchmark datasets, and results are compared with several state-of-the-art methods. The performance is evaluated on both quantitative and qualitative measures, and remarkable results are obtained.

3. Traditional anomaly detection methods localize anomaly in the scene containing a few tens of people. But in a dense crowd, what is happening is more important than who is doing it, therefore developed a global approach for crowd scene behavior classification (normal or abnormal) by minimizing the false alarm generation.

4. We validated our approach of crowd density estimation and crowd scene analysis on benchmarks datasets and obtained remarkable performance compared with existing methods.

6.3 Limitations and Future Work

There are many current research opportunities and new challenges open with this work for future action. The techniques implemented in this work for different tasks of crowd analysis can be used as basic building blocks for many real-world applications of crowd analysis. In the extension of the above research work done, some of the future research perspectives are as follows:

- Deep learning architecture can be used to process the massive amount of data generated by surveillance cameras.
- Potential enhancements incorporate the learning algorithms adapts to variable crowd density, and the texture of crowd should discriminate the non-ground plane regions, for instance, the texture of the sky can be confused with crowd textures.

- Localization of anomaly in the scene after anomalous scene detection will assist the security agents for better crowd monitoring.
- Crowd analysis tasks can be further extended for multiple camera scenarios.

Appendix A

List of Publications

International Journals

1. Sonu Lamba and Neeta Nain. A texture based mani-fold approach for crowd density estimation using gaussian markov random field. *Multimedia Tools and Applications*, Vol. 78(5), pp. 5645–5664, 2018.

2. Sonu Lamba and Neeta Nain. Segmentation of crowd flow by trajectory clustering in active contours. *Journal of The Visual Computer*, pp. 1–12 2019.

3. Sonu Lamba and Neeta Nain. Detecting anomalous crowd scenes by oriented tracklets' approach in active contour region. *Multimedia Tools and Applications*, Vol. 78(22), pp. 31101–31120, 2019.

International Conferences

1. Sonu Lamba, Neeta Nain, and Harendra Chahar. A robust multi-model approach for face detection in crowd. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2016 12th International Conference on*, pages 96–103. IEEE, Naples, Italy, 2016.

2. Sonu Lamba, Neeta Nain. Crowd Monitoring and Classification: A Survey. In *IC4S 2016, Springer Conference on*, Ajmer, India, 2016.

3. Sonu Lamba and Neeta Nain. Multi-source approach for crowd density estimation in still images. In *Identity, Security and Behavior Analysis (ISBA), 2017 IEEE International Conference on*, pages 1–6. IEEE, IIIT Delhi, India, 2017.

4. Poster entitled Multi-Source Approach for People Density Estimation in Crowded Scene, In Asia S&P 2017 for PhD Conclave held at SVNIT−Surat, India.

5. Sonu Lamba and Neeta Nain. A large scale crowd density classification using spatio-temporal local binary pattern. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2017 13th International Conference on*, pages 296–302. IEEE, Jaipur, India, 2017.

6. Sonu Lamba and Neeta Nain. Oriented Tracklets Approach for Anomalous Scene Detection in High Density Crowd. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2018 14th International Conference on*, pages 296–302. IEEE, Spain, 2018.

7. Sonu Lamba, Neeta Nain, A Literature Review on Crowd Scene Analysis and Monitoring. In *Pattern Recognition and Information Processing (PRIP'19), 2019* 14^{th} *International Conference on*, Belarus, 2019.

Bibliography

2006 umn dataset. `Unusualcrowdactivitydatasetofuniversityofminnesota`. online.

Ali, S. and Shah, M. (2007). A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE.

Ali, S. and Shah, M. (2008). Floor fields for tracking in high density crowd scenes. In *European conference on computer vision*, pages 1–14. Springer.

Bansal, A. and Venkatesh, K. (2015). People counting in high density crowds from still images. *arXiv preprint arXiv:1507.08445*.

Benabbas, Y., Ihaddadene, N., and Djeraba, C. (2011). Motion pattern extraction and event detection for automatic visual surveillance. *EURASIP Journal on Image and Video Processing*, 2011(1):163682.

Biswas, S., Praveen, R. G., and Babu, R. V. (2014). Super-pixel based crowd flow segmentation in h. 264 compressed videos. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 2319–2323. IEEE.

Brostow, G. J. and Cipolla, R. (2006). Unsupervised bayesian detection of independent motion in crowds. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 594–601. IEEE.

Brox, T. and Malik, J. (2010). Object segmentation by long term analysis of point trajectories. In *European conference on computer vision*, pages 282–295. Springer.

Caselles, V., Catté, F., Coll, T., and Dibos, F. (1993). A geometric model for active contours in image processing. *Numerische mathematik*, 66(1):1–31.

Chan, A. B., Liang, Z.-S. J., and Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer*

Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–7. IEEE.

Chan, A. B. and Vasconcelos, N. (2008). Modeling, clustering, and segmenting video with mixtures of dynamic textures.

Chan, A. B. and Vasconcelos, N. (2012). Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177.

Chan, T. F. and Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277.

Chang, T. and Kuo, C.-C. (1993). Texture analysis and classification with tree-structured wavelet transform. *IEEE Transactions on image processing*, 2(4):429–441.

Chen, D.-Y. and Huang, P.-C. (2013). Visual-based human crowds behavior analysis based on graph modeling and matching. *IEEE Sensors Journal*, 13(6):2129–2138.

Chen, K., Loy, C. C., Gong, S., and Xiang, T. (2012). Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3.

Cheriyadat, A. M. and Radke, R. J. (2008). Detecting dominant motions in dense crowds. *IEEE Journal of Selected Topics in Signal Processing*, 2(4):568–581.

Chetverikov, D. and Péteri, R. (2005). A brief survey of dynamic texture description and recognition. In *Computer Recognition Systems*, pages 17–26. Springer.

Cho, S.-Y., Chow, T. W., and Leung, C.-T. (1999). A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(4):535–541.

Choudri, S., Ferryman, J. M., and Badii, A. (2009). Robust background model for pixel based people counting using a single uncalibrated camera. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1–8. IEEE.

Cong, Y., Yuan, J., and Liu, J. (2013). Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851–1864.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Davies, A. C., Yin, J. H., and Velastin, S. A. (1995). Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, 7(1):37–47.

Dehghan, A. and Kalayeh, M. M. (2015). Understanding crowd collectivity: a meta-tracking approach. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–9. Citeseer.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. *International journal of computer vision*, 70(1):41–54.

Fradet, M., Robert, P., and Pérez, P. (2009). Clustering point trajectories with various life-spans. In *Visual Media Production, 2009. CVMP'09. Conference for*, pages 7–14. IEEE.

Frischholz, R. (2012). Bao face database at the face detection homepage.

Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., and Zhu, C. (2015). Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 43:81–88.

Gall, J., Yao, A., Razavi, N., Van Gool, L., and Lempitsky, V. (2011). Hough forests for object detection, tracking, and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 33(11):2188–2202.

Gao, Y., Liu, H., Sun, X., Wang, C., and Liu, Y. (2016). Violence detection using oriented violent flows. *Image and vision computing*, 48:37–41.

Gracia, I. S., Suarez, O. D., Garcia, G. B., and Kim, T.-K. (2015). Fast fight detection. *PLoS one*, 10(4):e0120448.

Gu, X., Cui, J., and Zhu, Q. (2014). Abnormal crowd behavior detection by using the particle entropy. *Optik-International Journal for Light and Electron Optics*, 125(14):3428–3433.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer.

Hassner, T., Itcher, Y., and Kliper-Gross, O. (2012). Violent flows: Real-time detection of violent crowd behavior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 1–6. IEEE.

Hou, Y.-L. and Pang, G. K. (2011). People counting and human detection in a challenging situation. *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, 41(1):24–33.

Hsu, R.-L., Abdel-Mottaleb, M., and Jain, A. K. (2002). Face detection in color images. *IEEE transactions on pattern analysis and machine intelligence*, 24(5):696–706.

Hu, M., Ali, S., and Shah, M. (2008). Learning motion patterns in crowded scenes using motion flow field. In *ICPR*, pages 1–5.

Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352.

Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., and Maybank, S. (2006). A system for learning statistical motion patterns. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1450–1464.

Hu, Y., Chang, H., Nian, F., Wang, Y., and Li, T. (2016). Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation*, 38:530–539.

Hussain, N., Yatim, H. S. M., Hussain, N. L., Yan, J. L. S., and Haron, F. (2011). Cdes: A pixel-based crowd density estimation system for masjid al-haram. *Safety Science*, 49(6):824–833.

Idrees, H., Saleemi, I., Seibert, C., and Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2547–2554. IEEE.

Jain, V. and Learned-Miller, E. (2010). Fddb: A benchmark for face detection in unconstrained settings. Technical report, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst.

Kang, D., Ma, Z., and Chan, A. B. (2018). Beyond counting: comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. *IEEE Transactions on Circuits and Systems for Video Technology*.

Kienzle, W., Franz, M. O., Schölkopf, B., and Bakir, G. H. (2005). Face detection—efficient and rank deficient. In *Advances in Neural Information Processing Systems*, pages 673–680.

Kong, D., Gray, D., and Tao, H. (2005). Counting pedestrians in crowds using viewpoint invariant training. In *BMVC*, volume 1, page 2. Citeseer.

Kratz, L. and Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models.

Krausz, B. and Bauckhage, C. (2012). Loveparade 2010: Automatic video analysis of a crowd disaster. *Computer Vision and Image Understanding*, 116(3):307–319.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kruthiventi, S. S. and Babu, R. V. (2015). Crowd flow segmentation in compressed domain using crf. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 3417–3421. IEEE.

Kuhn, A., Senst, T., Keller, I., Sikora, T., and Theisel, H. (2012). A lagrangian framework for video analytics. In *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*, pages 387–392. IEEE.

Laptev, I. and Lindeberg, T. (2003). Interest point detection and scale selection in space-time. In *International Conference on Scale-Space Theories in Computer Vision*, pages 372–387. Springer.

Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 878–885. IEEE.

Lempitsky, V. and Zisserman, A. (2010). Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332.

Li, J., Wang, T., and Zhang, Y. (2011). Face detection using surf cascade. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2183–2190. IEEE.

Li, M., Zhang, Z., Huang, K., and Tan, T. (2008). Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE.

Li, S. Z. (2009). *Markov random field modeling in image analysis*. Springer Science & Business Media.

Li, W., Mahadevan, V., and Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

Lim, M. K., Kok, V. J., Loy, C. C., and Chan, C. S. (2014). Crowd saliency detection via global similarity structure. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3957–3962. IEEE.

Lin, W., Mi, Y., Wang, W., Wu, J., Wang, J., and Mei, T. (2016). A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes. *IEEE Transactions on Image Processing*, 25(4):1674–1687.

Lloyd, K., Rosin, P. L., Marshall, D., and Moore, S. C. (2017). Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (glcm)-based texture measures. *Machine Vision and Applications*, 28(3-4):361–371.

Loy, C. C., Xiang, T., and Gong, S. (2009). Modelling multi-object activity by gaussian processes. In *BMVC*, pages 1–11. Citeseer.

Loy, C. C., Xiang, T., and Gong, S. (2012). Salient motion detection in crowded scenes. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*, pages 1–4. IEEE.

Lu, W.-C., Wang, Y.-C. F., and Chen, C.-S. (2010). Learning dense optical-flow trajectory patterns for video object extraction. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 315–322. IEEE.

Luvison, B., Chateau, T., Lapreste, J.-T., Sayd, P., and Pham, Q. C. (2011). Automatic detection of unexpected events in dense areas for videosurveillance applications. In *Video Surveillance*. InTech.

Ma, R., Li, L., Huang, W., and Tian, Q. (2004). On pixel count based crowd density estimation for visual surveillance. In *Cybernetics and Intelligent Systems, 2004 IEEE Conference on*, volume 1, pages 170–173. IEEE.

Ma, W., Huang, L., and Liu, C. (2010). Crowd density analysis using co-occurrence texture features. In *Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on*, pages 170–175. IEEE.

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE.

Marana, A., Costa, L. d. F., Lotufo, R., and Velastin, S. (1998). On the efficacy of texture analysis for crowd monitoring. In *Computer Graphics, Image Processing, and Vision, 1998. Proceedings. SIBGRAPI'98. International Symposium on*, pages 354–361. IEEE.

Marana, A. N., Costa, L. D. F., Lotufo, R., and Velastin, S. A. (1999). Estimating crowd density with minkowski fractal dimension. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 6, pages 3521–3524. IEEE.

Marsden, M., McGuinness, K., Little, S., and O'Connor, N. E. (2016). Fully convolutional crowd counting on highly congested scenes. *arXiv preprint arXiv:1612.00220*.

Mehran, R., Moore, B. E., and Shah, M. (2010). A streakline representation of flow in crowded scenes. In *European conference on computer vision*, pages 439–452. Springer.

Mehran, R., Oyama, A., and Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE.

Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, pages 69–82. Springer.

Milborrow, S., Morkel, J., and Nicolls, F. (2010). The muct landmarked face database. *Pattern Recognition Association of South Africa*, 201(0).

Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., and Murino, V. (2015). Analyzing tracklets for the detection of abnormal crowd behavior. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 148–155. IEEE.

Paragios, N. and Ramesh, V. (2001). A mrf-based approach for real-time subway monitoring. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE.

Pennisi, A., Bloisi, D. D., and Iocchi, L. (2016). Online real-time crowd behavior detection in video sequences. *Computer Vision and Image Understanding*, 144:166–176.

Rabaud, V. and Belongie, S. (2006). Counting crowded moving objects. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 705–711. IEEE.

Raghavendra, R., Del Bue, A., Cristani, M., and Murino, V. (2011a). Abnormal crowd behavior detection by social force optimization. In *International Workshop on Human Behavior Understanding*, pages 134–145. Springer.

Raghavendra, R., Del Bue, A., Cristani, M., and Murino, V. (2011b). Optimizing interaction force for global anomaly detection in crowded scenes. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 136–143. IEEE.

Rahmalan, H., Nixon, M. S., and Carter, J. N. (2006). On crowd density estimation for surveillance.

Rao, A. S., Gubbi, J., Marusic, S., and Palaniswami, M. (2016). Crowd event detection on optical flow manifolds. *IEEE transactions on cybernetics*, 46(7):1524–1537.

Rodriguez, M., Laptev, I., Sivic, J., and Audibert, J.-Y. (2011a). Density-aware person detection and tracking in crowds. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2423–2430. IEEE.

Rodriguez, M., Sivic, J., Laptev, I., and Audibert, J.-Y. (2011b). Data-driven crowd analysis in videos. In *ICCV 2011-13th International Conference on Computer Vision*, pages 1235–1242. IEEE.

Ryan, D., Denman, S., Fookes, C., and Sridharan, S. (2009). Crowd counting using multiple local features. In *2009 Digital Image Computing: Techniques and Applications*, pages 81–88. IEEE.

S. Wu, B. E. Moore, M. S. (2010). Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. *IEEE Conf. Comput. Vis. Pattern Recognit*, 43(6):2054–2060.

Sabzmeydani, P. and Mori, G. (2007). Detecting pedestrians by learning shapelet features. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Shao, J., Change Loy, C., and Wang, X. (2014). Scene-independent group profiling in crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2219–2226.

Sodemann, A. A., Ross, M. P., and Borghetti, B. J. (2012). A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1257–1272.

Solmaz, B., Moore, B. E., and Shah, M. (2012). Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):2064–2070.

Su, H., Yang, H., Zheng, S., Fan, Y., and Wei, S. (2013). The large-scale crowd behavior perception based on spatio-temporal viscous fluid field. *IEEE Transactions on Information Forensics and security*, 8(10):1575–1589.

Su, S. (2013). Boston marathon. flickr under creative commons attribution.

Subburaman, V. B. and Marcel, S. (2010). Fast bounding box estimation based face detection. In *ECCV, Workshop on Face Detection: Where we are, and what next?*, number EPFL-CONF-155015.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Thida, M., Eng, H.-L., and Remagnino, P. (2013). Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes. *IEEE Transactions on Cybernetics*, 43(6):2147–2156.

Thomaz, C. E. (2012). Fei face database. *FEI Face DatabaseAvailable*.

Tomasi, C. and Kanade, T. (1991). Detection and tracking of point features.

Tripathi, G., Singh, K., and Vishwakarma, D. K. (2019). Convolutional neural networks for crowd behaviour analysis: a survey. *The Visual Computer*, 35(5):753–776.

Vasilevskiy, A. and Siddiqi, K. (2002). Flux maximizing geometric flows. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1565–1578.

Vedaldi, A. and Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1469–1472. ACM.

Velastin, S., Yin, J., Davies, A., Vicencio-Silva, M., Allsop, R., and Penn, A. (1994). Automated measurement of crowd density and motion using image processing. In *Road Traffic Monitoring and Control, 1994., Seventh International Conference on*, pages 127–132. IET.

Vezhnevets, V., Sazonov, V., and Andreeva, A. (2003). A survey on pixel-based skin color detection techniques. In *Proc. Graphicon*, volume 3, pages 85–92. Moscow, Russia.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Viola, P., Jones, M. J., and Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161.

Wang, B., Ye, M., Li, X., Zhao, F., and Ding, J. (2012). Abnormal crowd behavior detection using high-frequency and spatio-temporal features. *Machine Vision and Applications*, 23(3):501–511.

Wang, C., Zhang, H., Yang, L., Liu, S., and Cao, X. (2015). Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1299–1302. ACM.

Wang, L. and Dong, M. (2012). Real-time detection of abnormal crowd behavior using a matrix approximation-based approach. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2701–2704. IEEE.

Wang, X., Ma, X., and Grimson, E. (2007). Unsupervised activity perception by hierarchical bayesian models. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Wang, X., Ma, X., and Grimson, W. E. L. (2009). Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on pattern analysis and machine intelligence*, 31(3):539–555.

Wang, Y.-Q. (2014). An analysis of the viola-jones face detection algorithm. *Image Processing On Line*, 4:128–148.

Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 90–97. IEEE.

Wu, B. and Nevatia, R. (2007). Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266.

Wu, S. and San Wong, H. (2012). Crowd motion partitioning in a scattered motion field. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(5):1443–1454.

Wu, S., Wong, H.-S., and Yu, Z. (2014). A bayesian model for crowd escape behavior detection. *IEEE transactions on circuits and systems for video technology*, 24(1):85–98.

Wu, S., Yu, Z., and Wong, H.-S. (2009a). Crowd flow segmentation using a novel region growing scheme. In *Pacific-Rim Conference on Multimedia*, pages 898–907. Springer.

Wu, S., Yu, Z., and Wong, H.-S. (2009b). A shape derivative based approach for crowd flow segmentation. In *Asian Conference on Computer Vision*, pages 93–102. Springer.

Wu, Y., Wang, Y., and Jia, Y. (2013). Adaptive diffusion flow active contours for image segmentation. *Computer Vision and Image Understanding*, 117(10):1421–1435.

Xiaohua, L., Lansun, S., and Huanqin, L. (2006). Estimation of crowd density based on wavelet and support vector machine. *Transactions of the Institute of Measurement and Control*, 28(3):299–308.

Xiong, G., Wu, X., Chen, Y.-L., and Ou, Y. (2011). Abnormal crowd behavior detection based on the energy model. In *Information and Automation (ICIA), 2011 IEEE International Conference on*, pages 495–500. IEEE.

Xu, J., Denman, S., Sridharan, S., Fookes, C., and Rana, R. (2011). Dynamic texture reconstruction from sparse codes for unusual event detection in crowded scenes. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 25–30. ACM.

Xu, L., Gong, C., Yang, J., Wu, Q., and Yao, L. (2014). Violent video detection based on mosift feature and sparse coding. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3538–3542. IEEE.

Yadav, S. and Nain, N. (2015). Fast face detection based on skin segmentation and facial features. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2015 11th International Conference on*, pages 663–668. IEEE.

Yadav, S. and Nain, N. (2016). A novel approach for face detection using hybrid skin color model. *Journal of Reliable Intelligent Environments*, 2(3):145–158.

Yang, H., Cao, Y., Wu, S., Lin, W., Zheng, S., and Yu, Z. (2012). Abnormal crowd behavior detection based on local pressure model. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE.

Yuan, Y., Fang, J., and Wang, Q. (2015). Online anomaly detection in crowd scenes via structure analysis. *IEEE Transactions on Cybernetics*, 45(3):548–561.

Zhang, C., Li, H., Wang, X., and Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 833–841. IEEE.

Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597.

Zhao, J., Xu, Y., Yang, X., and Yan, Q. (2011). Crowd instability analysis using velocity-field based social force model. In *Visual Communications and Image Processing (VCIP), 2011 IEEE*, pages 1–4. IEEE.

Zhou, B., Tang, X., and Wang, X. (2012a). Coherent filtering: detecting coherent motions from crowd clutters. In *Computer Vision–ECCV 2012*, pages 857–871. Springer.

Zhou, B., Tang, X., and Wang, X. (2013). Measuring crowd collectiveness. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3049–3056.

Zhou, B., Wang, X., and Tang, X. (2012b). Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2871–2878. IEEE.