VISUAL ANALYSIS OF DENSE CROWDS

KOK VEN JYN

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2016

VISUAL ANALYSIS OF DENSE CROWDS

KOK VEN JYN

THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2016

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: KOK VEN JYN

Registration/Matrix No.: WHA120015

Name of Degree: Doctor of Philosophy

Title of Thesis: Visual Analysis of Dense Crowds

Field of Study: Computer Vision (Computer Science)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRACT

The steady worldwide population growth with continuing urbanization renders the formation of crowd by chance a norm. The mere existence of crowd has the prospect of progressing into a hazardous scene. Consequently, visual analysis of dense crowds is a growing research topic in the domain of computer vision. Conventional visual analysis methods are mostly object-centric, thus, are neither suitable nor capable of analyzing dense crowd. Hence, this thesis proposes novel solutions to analyze images and videos of dense crowds, which contain hundreds to thousands of individuals. The main objective are, first, to obviate the difficulty of segregating individuals in dense crowd scenes to infer dense crowd segments, secondly to estimate the number of individuals and finally to detect unusual events, by exploiting spatial and temporal cues readily available from the scenes.

Dense crowd segmentation generally serves as one of the essential steps for further visual analysis of the dense crowds. The thesis first demonstrates the significance of simplifying dense crowd scenes into structurally meaningful atomic regions for dense crowd segmentation. This proposed approach is formulated using the concept and principles of granular computing. It shows that by exploiting the correlation among pixel granules, structurally similar pixels can be aggregated into meaningful atomic structure granules. This is useful in outlining natural boundaries between crowd and background (i.e. non-crowd) regions necessary for dense crowd segmentation. Moreover, the proposed approach is scene-independent; thus it can be applied effectively to dense crowd scenes with a variety of physical layout and crowdedness.

Second, this thesis presents an approach to utilize irregular patches conforming to the natural outline between crowd and background to estimate the number of individuals in dense crowd scenes. As opposed to most of the existing approaches that uses pixel-grid

representation, the proposed density estimation approach allows a model to adapt itself to the arbitrary distribution of crowd where the underlying spatial information of scenes can be accurately extracted. Here, a direct mapping is established between the extracted features and the number of people.

Third, to detect saliency in dense crowd scenes, low-level features extracted from the crowd motion field are transformed into a global similarity structure. This global similarity structure representation allows the discovery of the intrinsic manifold of the motion dynamics, which could not be captured by the low-level representation. Most importantly, unlike conventional methods, the proposed approach does not require *tracking,* and *prior information or model learning* to identify interesting / salient regions in the dense crowd scenes.

These proposed approaches are validated by using public dataset of dense crowd scenes. From the empirical results, it is anticipated that the collective analysis of this thesis will constitute a complete dense crowd analysis system that is able to infer regions of dense crowds, estimate crowd density and identify saliency in mass gathering for proactive crowd management.

ABSTRAK

Perhimpunan orang ramai di tempat awam secara tidak sengaja menjadi suatu perkara yang norma akibat pertambahan populasi penduduk dunia dan perkembangan perbandaran yang berterusan. Kerumunan orang ramai berpotensi untuk berubah menjadi senario yang berbahaya, seperti rempuhan orang ramai. Sehubungan itu, analisis visual himpunan orang ramai merupakan satu topik penyelidikan yang semakin berkembang dan giat diterokai dalam domain visi komputer. Kebanyakan kaedah analisis visual yang konvensional memfokuskan objek. Oleh itu, kaedah tersebut tidak sesuai dan tidak mampu untuk menganalisi himpunan orang ramai. Justeru, tesis ini mencadangkan penyelesaian yang baharu untuk menganalisis gambar-gambar dan video yang mengandungi ratusan hingga ribuan orang ramai. Objektif utama tesis ini adalah, pertamanya, untuk menangani kesulitan dalam usaha mengasingkan individu daripada himpunan orang ramai bagi menentukan segmen orang ramai daripada segmen latar belakang, keduanya, untuk menganggarkan bilangan individu dan akhirnya, mengesan perkara yang menonjol (*salient*), dengan mengunakan maklumat berkenaan *spatial* dan *temporal* yang didapati daripada gambar-gambar dan video tentang orang ramai.

Segmentasi himpunan orang ramai yang padat (*Dense crowd segmentation*) pada umumnya berfungsi sebagai salah satu langkah penting untuk analisis visual orang ramai yang selanjutnya. Tesis ini pada awalnya menunjukkan kepentingan membahagikan himpunan orang ramai kepada kelompok kecil yang bermakna untuk *dense crowd segmentation*. Pendekatan yang diusulkan ini digubah dengan menggunakan konsep dan prinsip-prinsip pengkomputeraan granul (*granular computing*). Hal ini menunjukkan bahawa dengan mengeksploitasi hubungan antara granul piksel, struktur piksel yang sama dapat digabungkan untuk menjadi struktur granul yang bermakna. Pendekatan ini berguna dalam merangka sempadan semula jadi antara kumpulan orang ramai dan latar belakang

yang diperlukan untuk *dense crowd segmentation*. Tambahan pula, pendekatan yang dicadangkan ini boleh digunakan dengan berkesan bagi himpunan orang ramai dalam pelbagai persekitaran dan kesesakan.

Kedua, tesis in menyampaikan satu pendekatan untuk menggunakan kelompok tidak sekata yang mematuhi sempadan semula jadi antara kumpulan orang ramai dan latar belakang bagi menganggarkan bilangan individual. Berbeza dengan kebanyakan pendekatan sedia ada yang menggunakan grid piksel, pendekatan anggaran kepadatan yang dicadangkan membolehkan algorithma menyesuaikan dirinya dengan sempadan kawasan orang ramai. Dalam pada itu, maklumat *spatial* yang sedia ada dalam himpunan dapat diekstrak dengan tepat.

Ketiga, untuk mengesan ketonjolan (*saliency*) dalam himpunan orang ramai, *low-level feature* yang diekstrak daripada pergerakan orang ramai diubah menjadi *global similarity structure*. *Global similarity structure* membolehkan penemuan manifold intrinsik dalam dinamik gerakan yang tidak dapat dikesan oleh *low-level representation*. Yang pentingnya, berbeza dengan kaedah konvensional, pendekatan yang dicadangkan tidak memerlukan pengesanan (*tracking*) dan maklumat terdahulu (*prior information*) atau pembelajaran model (*model learning*) untuk mengenal pasti kawasan-kawasan yang menonjol (*salient*) dalam himpunan orang ramai.

Pendekatan-pendekatan yang dicadangkan disahkan dengan menggunakan set data awam himpunan orang ramai. Daripada keputusan empirikal, analisis kolektif tesis ini dijangkakan akan menjadi sistem analisis himpunan orang ramai yang lengkap dan mampu menentukan segmen orang ramai, menganggarkan kepadatan orang ramai dan mengenal pasti ketonjolan (*saliency*) dalam perhimpunan besar-besaran bagi pengurusan orang ramai secara proaktif.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my supervisor, Dr. Chan Chee Seng, who has given me the opportunity to pursue my doctoral studies. I am deeply grateful for your guidance, patience, support and constant encouragement throughout the period of my research and preparation of this thesis. You have provided me and other lab members sufficient room and time, as well as a conducive research environment, which encourages us to be independent in conducting original researches.

Besides, I would also like to thank Dr. Loy Chen Change, who challenges and pushes me to improve and be a better researcher. I am deeply grateful for your invaluable advice and guidance, which is instrumental in shaping my approach to this research. My warm appreciation also goes to various members and staff at Multimedia Laboratory in Chinese University of Hong Kong.

Additionally, I would like to thank my fellow research colleagues for your friendship, as well as making the lab a fun and productive environment. I am especially grateful to Dr. Vembarasan Vaitheeswaran for your assistance in mathematical formulations and for always came through when I needed help in my research.

Last but not least, I wish to express my deepest love and gratitude to my family. Words are too pale to express how grateful I am to my parents, sisters and parents-in-law for your unconditional love, care, support, encouragement and understanding. Thank you mom and dad for showing faith in me and giving me liberty to choose what I desired. My love and thanks to my beloved husband and best friend, Loh Jianwei, my constant companion, who always stood by my side supporting and motivating me all the way through till the end of my PhD journey. Thank you for your sacrifice and embracing my dreams as you do your own.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AD	:	Absolute Difference.
CCTV	:	closed circuit television camera.
DSIFT	:	Dense Scale-Invariant Feature Transform.
GrC	:	granular computing.
GrCS	:	granular computing based dense crowd segmentation.
kNN	:	k-nearest neighbors.
KRR	:	Kernel Ridge Regression.
LBP	:	Local Binary Pattern.
LRI	:	Local Range of Intensity.
LSD	:	Local Standard Deviation.
MRF	:	Markov Random Field.
PC	:	Phase Congruency.
RBF	:	Radial Basis Function.
RF	:	Random Forest.

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

In mid-2013, the world population reached 7.2 billion, 648 million more than in 2005 (United Nations, Department of Economic and Social Affairs, Population Division, 2013). According to the population estimates and projections from the United Nations, Department of Economic and Social Affairs, Population Division (2013), the world population is expected to reach 9.6 billion by the year of 2050. Globally, over half (54 percent) of the world population resides in urban areas in 2014 and is expected to increase to 66 percent by 2050 (United Nations, Department of Economic and Social Affairs, Population Division, 2014). The worldwide population growth, coupled with the continuing urbanization has rendered the occurrence of crowded environment a growing norm. The presents of large crowd in any environment can disrupts and challenges the effectiveness of public management, safety and security. It is therefore not surprising that computer vision researchers have become increasingly focused on visual crowd analysis (Ali, Nishino, Manocha, & Shah, 2013). Substantial efforts have been made toward understanding crowded scenes for crowd analysis in both static images and video sequences. This endeavor is motivated by the need of a sophisticated crowd surveillance system. A significant application of computer vision based visual crowd analysis is intelligent crowd surveillance (Rodriguez, Sivic, & Laptev, 2012), which aims to automatically infer crowds segments for density estimation and subsequently detect unusual events that could pose a significant threat to public safety and security in crowded environments.

1.1 Visual Analysis of Dense Crowds

A mass gathering of individuals, i.e. crowding, can be either planned well in advance (e.g. concert, parade, festival, rally and religious event) or take place spontaneously (e.g. crowd in a train station during rush hour). The individuals in the crowd, in the most

basic sense, gather at a specific venue with a coherent purpose for a length of time (World Health Organization (WHO), 2008), where the behavior of one individual is influenced by the other.

Due to the large number of individuals in close proximity, any mass gathering is at high risk of turning fatal given physical stress (i.e. overcrowding) or sudden external stress (e.g. shooting, fire, terrorist attack) (Helbing, Johansson, & Al-Abideen, 2007). This is evident with the recurrent of lethal crowd disasters. Figure 1.1 illustrates sample images captured during the progression of some high-profile crowd disasters happened globally. Thus, visual surveillance of mass gathering in public settings is commonplace, predominantly in response to the dynamic and degenerating risk to public safety and security (Moore, Ali, Mehran, & Shah, 2011).

(a) Hillsborough disaster 1989 (Taylor, 1990)

(b) Love Parade disaster 2010 (Helbing & Mukerji, 2012)

(c) Shanghai New Years Eve disaster 2014 (Kaiman, 2015)

Figure 1.1: Sample images captured during the progression of some crowd disasters. (a) Hillsborough disaster: claimed 66 innocent lives and injured 140. (b) Love Parade disaster: claimed 21 innocent lives and injured 510. (c) Shanghai New Years Eve disaster: claimed 36 innocent lives.

Conventional visual surveillance systems depend heavily on human operators to operate and monitor a set of television screens to intercept trouble before it occurs and followed by determining the next course of action upon occurrence of an incident. The effectiveness and efficiency of such surveillance systems are subjected to the vigilance of the operators. However, visual surveillance task at mass gatherings where crowd of hundred or thousand gathers is substantially more taxing compared to scenes with a few

number of people, primarily due to the extent of activity occurred within such scenes. In fact, even the most diligent human operators would face substantial challenges in performing basic visual recognition task, such as counting individuals in mass gatherings to predict overcrowding crisis. This perhaps makes scenes with dense crowd in dire need of intelligent surveillance.

Furthermore, in addition to the common causes of overlook during surveillance (i.e. short attention span, fatigue due to prolonged monitoring and excessive amount of television screens to monitor), crowd scenes also present a new set of challenges. Visual crowd surveillance is challenging due to (1) the sheer number of individuals in scenes, and (2) severe occlusions between individuals. With the increase of individuals in a scene, it would demand a greater effort during the visual inspection. Denser crowd would, furthermore, amplify the likelihood of occlusions, making it difficult to discern each individual. This can compromise one's capability to monitor and focus the attention on anomalies of any scales, while ignoring the clutters in a crowded scene. The explanation has to do with a *pop-out effect* (Szeliski, 2010), where increasing distractors (i.e. individuals, in the context of this thesis) would hinder the parallel processing to pinpoint any unusual events and scrutinize the desired individual.

Technology and service providers (e.g. CrowdVision, NEC, and AGT International) as well as end-users have recognized that manual surveillances of dense crowd scenes alone is inadequate to meet the sought after level of accuracy and precision in crowd surveillance systems. Several tragic crowd disasters from the past (see Table 1.1) have implied the significance of an intelligent visual crowd surveillance system for a proactive crowd management to anticipate disaster and provide support in good time. An intelligent crowd surveillance system is paramount to minimize deleterious impact of dense crowd under adverse conditions. To fulfill such a need, advance computer vision techniques (Junior, Musse, & Jung, 2010) are relentlessly being explored and incorporated into visual

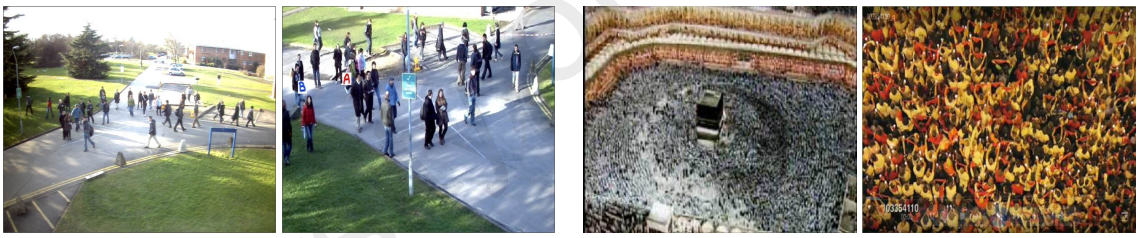
Table 1.1: Examples of crowd disasters at mass gatherings.

Date	Event - Place	Description	Casualties	Reference
1971	Football match Glasgow, UK	Crush between fans entering and exiting.	66 deaths 140 injured	(Popplewell, 1986)
1981	Nightclub fire Dublin, Ireland	Fire was started deliberately in the alcove.	48 deaths 128 injured	(Tribunal of Inquiry on the Fire at the Stardust, Artane, Dublin, 1981)
1989	Football match Sheffield, UK	Crush due to overcrowding surge against barriers.	96 deaths 766 injured	(Taylor, 1990)
1990	The Hajj Mecca, Saudi Arabia	Stampede due to overcrowding in a pedestrian tunnel leading out from Mecca.	1426 deaths	(Ahmed, Arabi, & Memish, 2006)
1991	Football match Orkney, South Africa	Panicking fans try to escape from brawls that break out in the grandstand.	42 deaths 50 injured	(Darby, Johnes, & Mellor, 2005)
1993	New Year's Eve revelry Hong Kong	Revelers fell and pill on top of another when rushing down a steep cobblestone walkway wet with beer and party foam.	21 deaths	(Bokhary, 1993)
1994	The Hajj Mecca, Saudi Arabia	Progressive crowd collapse as a result of overcrowding of pilgrims.	270 deaths	(Gad-el Hak, 2008)
1995	School's annual function Mandi Dabwali, India	Stampede due to panicking crowd tried to escape a sudden fire.	441 deaths 150 injured	(Moddie, 2004)
2000	Football match Harare, Zimbabwe	Stampede broke out as fans rushing to get away from the noxious fumes of tear gas fired by the police.	13 deaths	(Madzimbamuto, 2003)
2004	Miyun lantern festival Beijing, China	Crush due to overcrowding on a bridge.	37 deaths 24 injured	(Zhen, Mao, & Yuan, 2008)
2006	Philsports stadium Manilla, Philippines	Individuals at the front of the crowd stumbled, which lead to a dominoes effect and stampede.	74 deaths 627 injured	(Lee, 2012)
2008	Ramadan alms giving Java, Indonesia	Crush due to crowd surging forward to fight over alms (i.e. zakat) handed out as a Ramadan gift.	23 death	(MacKinnon, 2008)
2010	Love Parade music festival - Duisburg, Germany	Crush due to overcrowding at a narrow tunnel leading into the festival.	21 death 510 injured	(Helbing & Mukerji, 2012)
2010	Khmer water festival Phnom Penh, Cambodia	Bottleneck on the bridge has triggered sudden panic in the crowd, which lead to stampede.	347 death 755 injured	(Hsu & Burkle, 2012)
2013	Boston marathon Massachusetts, USA	Two pressure cookers explored near where the crowd gathers.	3 death 264 injured	(Starbird, Maddock, Orand, Achterman, & Mason, 2014)
2014	New Year's Eve revelry Shanghai, China	Crush between crowd climbing up and down a stairway.	36 death 42 injured	(Kaiman, 2015)

crowd surveillance systems to assist human operators in surveillance tasks.

1.1.1 Dense Crowd Analysis in Computer Vision

In crowd scenes where individuals are distinguishable (see Figure 1.2a), analysis of each individual (e.g. tracking) may be possible (Idrees, Warner, & Shah, 2014). However, when the size of a collection of entity increases to an extent that one's bodily movement can no longer be distinguished and with only a few pixels per individual (see Figure 1.2b), human actions become group activity and eventually crowd behavior. For example, in a crowded theater, audiences often move in a synchronize pattern during ingress and egress from the venue. Visual analysis of each individual become less feasible and less relevant, whilst understanding the crowd as a whole for an enhanced visual analysis is of more interest.



(a) Sparse crowds **(b)** Dense crowds

Figure 1.2: Examples of crowd scenes: (a) sparse crowds where individuals are distinguishable, and (b) dense crowds where there are only few pixels per individual. The primary interest of this thesis is the analysis of dense crowd scenes.

Crowd analysis is a growing research topic in the field of computer vision fueled by the need to carry out visual surveillance in dense crowd scenes (Zhan, Monekosso, Remagnino, Velastin, & Xu, 2008; Ali et al., 2013). It is an integral part of a wide array of applications with direct social impact, i.e. entertainment industry (e.g. animation of crowd in movies and games), advertising industry, as well as public safety and security. For instance, the computerize processing and analysis of crowd can support and assist human operators by highlighting circumstances which require closer examination. This changes the role of human operator from an observer to overseer to alleviate the likeli-

hood of important incidents left unnoticed during surveillance. Pre-recorded imagery of crowd scenes can also be analyzed to extract crowd and event information for the purpose of post-event forensic investigation or crowd simulation. It serves as an effective tool to establish global situational awareness. In retail and hospitality industries, crowd analysis can be an intelligence gathering tool (Tian et al., 2008) which provide valuable information to evaluate retail performance across multiple locations at different times of the day. The gathered information can also be used to improve and optimize customer service, floor plan and advertising program (Loy, Chen, Gong, & Xiang, 2013).

Despite the significant advancement of computer vision research, particularly, human motion analysis (J. K. Aggarwal & Cai, 1999), most of the existing work has been focused on individuals or small group of individuals in non-crowded scenes. Conventional visual analysis methods are mostly object-centric where they learn the behavior of the scene in three steps: object detection, tracking and compilation of tracked results for individuals or global behavior modelling (Ali et al., 2013). The applicability of such approach is limited to scenes with relatively few individuals (approximately 5-20 individuals) (Ali et al., 2013). This is because it is difficult to discern individuals in dense crowd since they are in close proximity with each other (Rodriguez et al., 2012). Similarly, as noted by Zhan et al. (2008), conventional computer vision methods work well on sparse scenes, but are inadequate to analyze crowded scenes. Correspondingly, a straightforward extension of these methods is neither suitable nor capable to analyze dense crowd. This is because a crowd is beyond a simple sum of individuals, where it can assume different complex behaviors. The difficulty of analyzing crowd increases disproportionately in relation to the number of individuals in a crowd. Under such circumstances, dense crowd analysis is a unique research problem which needs to be specifically addressed. There has been a series of research studies in crowd analysis at macroscopic and/or microscopic level. The microscopic level deals with the crowd as discrete individuals while the macro-

scopic level treats the crowd as a unit. For dense crowd analysis, where analysing each individual is difficult, the contextual (spatial) and temporal constraints can be employed to analyse scenes at macroscopic level. Holistic properties of dense crowd scenes are usually extracted to build crowd motion model (Ali & Shah, 2007; Mehran, Moore, & Shah, 2010; Wu et al., 2009).

In terms of experimental data, most of the available datasets have low to medium density crowd. For instance, Mall dataset has a density of 13-53 individuals per frame (K. Chen, Loy, Gong, & Xiang, 2012), PETS dataset has a density of 3-40 individuals per frame (Ferryman & Ellis, 2010) and USCD dataset contains 11-46 individuals per frame (Chan, Liang, & Vasconcelos, 2008). Only in the recent years with the rising of dense crowd analysis in the field of computer vision, more dense crowd datasets are being introduced for evaluation, such as, the UCF crowd counting dataset (Idrees et al., 2013). Images in the UCF crowd counting dataset contain between 94 and 4545 people per image, with an average of 1280 people over 50 images. Such high density crowd scenes imply that there are only a few pixels per individual, thereby exacerbating visual analysis of individuals in crowd.

To achieve dense crowd analytics in surveillance system demands more sophisticated computer vision algorithms exclusive to dense crowds. Various computer vision techniques (Idrees et al., 2014; Kang & Wang, 2014; Shao et al., 2015; Cao, Zhang, Ren, & Huang, 2015) were being explored recently to make better use of contextual (spatial) and / or temporal information for crowd analysis. In this thesis, the research is greatly motivated by the need to have an enhanced computer vision system to analyze dense crowd, with the aim to improve crowd safety and security. This thesis explores the use of contextual (spatial) information to *infer crowd segments* and *estimate crowd density*. Subsequently, temporal information is used in order to *detect unusual events* in dense crowd scenes.

(a) A Colony of ants. **(b)** A flock of starlings. **(c)** A crowd of people.

Figure 1.3: A variety of entities that made up a crowd in nature, such as (a) ants, (b) birds and (c) people.

1.1.2 Dense Crowd - Definition

According to the Oxford English Dictionary[1], '*crowd*' is a generic term that refers to a large number of entities gathered together, and the term '*dense*' refer to the condition of having each constituent entity closely compacted together. In this thesis, a clear distinction is made between 'crowd' and 'dense crowd'. The term 'crowd' has been broadly used in the computer vision community when referring to a collection of entities of varying crowdedness. The impact of crowdedness is important to understand for crowd safety (Still, 2000). To reduce the ambiguities, the term 'dense crowd' will be used exclusively to describe a large number of densely packed entities, as shown in Figure 1.2b and Figure 1.3c, and it will be used consistently throughout the remainder of the thesis.

Entities in crowd can be of a variety of types (as shown in Figure 1.3) including but not limited to people, vehicle, fish, bird, ant and bacteria. Various researches have been conducted focusing on different collection of entities (e.g. ant colonies (Deneubourg, Pasteels, & Verhaeghe, 1983), fish schools (Kunz & Hemelrijk, 2003) and bird flocks (Heppner & Grenander, 1990). In this thesis, the primary interest is the analysis of dense crowd scenes that contain crowd of people, with the aim to enhance crowd safety and security. According to Dubos (1974), the research of human crowd is more complex than the animal population given that human is profoundly conditioned by social and cultural

[1]Oxford English Dictionary: http://www.oed.com/

determinants. For instance, the appropriate social distance between individuals in crowd varies from culture to culture. Some people may require smaller space whereas some people may demand greater physical separation.

1.1.3 What is Unusual Event in Dense Crowd?

The notion of *unusual event* has been referred to with various terms in different studies. The different terms in use includes abnormal, interesting, irregular, suspicious event, or simply saliency, anomaly, and outlier. These terms may refer to event, behavior or activity, thus causing much confusion in the literature.

Throughout the literature, *unusual event* is treated as a context-sensitive notion. That is, these terms are used in various studies to define or distinguish the notion according to the study of interest. For example, Loy et al. (2012) define unusual events as salient motion in crowded scenes when the motion flows deviate from regular instances. Analogously, unusual events are referred to as anomalous or abnormal events that are dissimilar from the normal crowd behavior by Mahadevan, Li, Bhalodia, and Vasconcelos (2010). W. x. Li, Mahadevan, and Vasconcelos (2014)'s definition of unusual event is those that have low probability with respect to the probabilistic model of normal behavior. Kratz and Nishino (2009) identify unusual event as statistical deviations of the same scene. Unusual events may also denote circumstances such as overcrowding (Chiappino, Morerio, Marcenaro, & Regazzoni, 2014).

Despite the fact that no consensus had been established in the literature, this thesis focuses on the commonalities of the notion. Specifically, for visual analysis of dense crowd, an event is deemed worthy of being highlighted if the event is unknown, unpredictable or has not been learned before (i.e. low statistical representation in a dataset). The researcher of this thesis made no distinction between the various terms used to denote unusual events. Rather, the terms are used interchangeably throughout the remainder

Figure 1.4: Dense crowd observed in real world environment. Note that the crowd in different scenes exhibit drastic appearance variations due to illumination conditions, inter-occlusions, camera orientations and pose changes. Best viewed in color.

of this thesis.

1.1.4 Objectives of Study

The principle goal of this study is to devise computer vision algorithms for *dense crowd analysis*. Specifically, the work focuses on three research associated with dense crowd analysis: (1) localization of crowd segments in public scenes by obviating the difficulties of segregating individuals, (2) crowd density estimation in public scenes using irregular patches conforming to the natural outline between crowd and background regions, and (3) detection of unusual events in crowded public scene, to assist human in improving dense crowd safety and security.

In the following section, the underlying challenges faced by the research community and the problem formulation are discussed, which serve as the main motivation of this study to achieve the research aims and objectives.

1.2 Challenges and Problem Formulation

Dense crowd scenes pose distinctive challenges that severely impede the development of robust visual analysis methods for intelligent crowd surveillance. The main aspects which make dense crowd scenes intrinsically difficult to analyze is due to several inextricable factors as follows:

1. **Choice of Granularity** – Dense human crowd is complex as it exhibits large dynamic and psychological characteristic variation which are often dependent on the situations and environmental settings. It could also be associated with the characteristics of each independent individual such as age, sex and cultural background (Ali et al., 2013). This makes it challenging to determine the optimal size of granularity (e.g. pixel-, patch-, individual- or image-based analysis) to analyze dense crowd scenes (Ali, 2008). Thus, granules that can be adaptive to dense crowd segments are essential for dense crowd analysis.

2. **Appearance Variations of Crowd** – Crowd across all scenes varies drastically because of different crowdedness, illumination conditions, inter-occlusions and variations of clothing and poses (see Figure 1.4). At the same time, perspective distortions due to camera orientation and position implicate changes of scales of individuals within a crowd. Moreover, crowds tend to be heterogeneous in nature, where different portions of the crowd within the same environment could behave contrastively. This entails exploring other information from crowd imagery to disambiguate appearance information.

3. **Few Pixels per Individual** – It is infeasible to discern individuals and one's body parts due to low resolution imagery, where an individual may only be occupying a few pixels per individual, as shown in Figure 1.4. Hence, applying conventional object-centric strategy that requires explicit person detection and tracking in dense crowd are still in their infancy stage (Idrees, Soomro, & Shah, 2015).

4. **Effects of Terrain & Scene Features** – The formation of crowd across different scenes is inherently dependent on the constraints imposed by the environmental layout. Behavior of individuals in crowd can vary drastically based on the given situations and layouts. Moreover, one can observe in Figure 1.4 that background

regions that consist of trees, buildings, vehicles and carpet grasses may clutter in such a way that it resembles crowd regions. This hampers the process to obtain a good separation between crowd and background regions for crowd analysis. Nevertheless, the scene texture can be used as cues to differentiate between crowd and background regions.

5. **Representation of Abnormality in Crowd** – The definition of interesting region in crowd has been causing much debate in the literature due to the subjective nature and complexity of the human behaviors. Some researchers consider any deviation from the ordinary observed events as anomaly, whereas others consider rare or outstanding event as interesting. One may question the benefit of predefining the various types of anomalies in dense crowds, in serving as the cue to anticipate crowd disasters.

1.2.1 Localization of Dense Crowd Segments in Public Scenes

Generally in crowd analysis, crowd segmentation serves as one of the fundamental steps for further analysis, such as crowd density estimation (Idrees et al., 2013) and crowd behavior analysis (Solmaz et al., 2012). This is also stated in (Idrees et al., 2015) and (Kang & Wang, 2014) that the localization of crowd segments is required prior to visual tasks such as tracking or behavior understanding.

Nevertheless, due to the aforementioned *factors 1– 4*, inferring crowd segments in dense crowd scenes is taxing. This motivates the use of contextual (i.e. spatial) information to decompose crowd scene image into different levels of granularity to avoid actual segregation of individuals. This is similar to human cognition in problem solving. In essence, the dichotomy articulated by Moravec (Moravec, 1988) between humans and machines regarding the easiness and complexity in solving different problems remains valid today. Specifically, machines perform poorly in tasks that are seemingly effortless

Figure 1.5: Example of a crowd image divided into segments. Green outline indicates the partitions between segments. (Red bounding boxes) segments consisting of crowd and background (non-crowd) regions. Best viewed in color.

and natural for humans (i.e. recognizing crowd regions), but can easily solve problems that humans find challenging (i.e. numerical computation). One key advantage of the human mind has over a machine in cognition is the ability to segment visual information into meaningful units of analysis effortlessly (Hendee & Wells, 1997).

Thus, a crowd segmentation strategy that simulates human cognitive process would be extremely useful to achieve abstraction on the essential details at different granularities for effective crowd segmentation. In fact, studying the correlation among image granules at different levels of granularity is required to simplify image scene into meaningful atomic regions that adhere to the natural boundaries between crowd and non-crowd regions. At present, this problem is nontrivial and has not been addressed before for dense crowd segmentation in public scenes.

1.2.2 Density Estimation in Dense Crowd Scenes

Visual crowd analysis for density estimation in public scenes can be a highly effective means to ensure public safety and security. However, estimating the density of individuals in a dense crowd scene is intrinsically difficult. Specifically, the aforementioned *factors 2 – 3* render a direct implementation of conventional object centric strategy (i.e. object detection and tracking) infeasible. The problem is further hampered by the ambiguities caused by varying physical layout of crowd environments (i.e. *factor 4*).

To overcome the complexity of density estimation in the dense crowd scenes, most

methods (Idrees et al., 2013; Davies, Yin, & Velastin, 1995) employ a regression strategy, in which a model is trained to map the correlation between the holistic and collective description of crowd patterns to the number of individuals. Additional measures could be taken to alleviate perspective distortion by dividing the image space into smaller segments (i.e. pixel-grid). However, the regression strategy is not feasible when it is used in unconstrained public scenes, where there may be cases that only partial of a segment consist of crowd (as illustrated in Figure 1.5). This can lead to inaccurate description of crowd patterns for density estimation.

Thus, a method is required to be adaptive to varying physical layout of different crowd scenes and at the same time able to extract the most critical and discriminative descriptions of crowd patterns for an enhanced density estimation in dense crowd scenes.

1.2.3 Dense Crowd Saliency Detection

Besides the aforementioned *factor 1– 4*, one of the foremost challenges in saliency detection in dense crowd scenes is the representation of crowd abnormality (i.e. *factor 5*). The major implication is that collecting sufficient training data which addresses each possible abnormal scenario in the dense crowd scenes for supervised learning will be impractical. This is because human behavior is extremely complex, diverse, changing and unusual events are unpredictable. Consequently, most crowd saliency detection methods (Rodriguez, Sivic, Laptev, & Audibert, 2011; B. Zhou, Wang, & Tang, 2012) that commence by learning an activity model of the scene, followed by using the learned model to detect anomalies may be limited to the detection of the learned behaviour. They are not adaptive to diverse deployment scenarios.

To cope with crowd saliency variations, a method that alleviates the need for a learned model and at the same time requires no segregations of individuals in crowd and prior information is essential to detect crowd saliency.

1.3 Contributions

The contributions of this thesis to visual analysis of dense crowd, particularly on localizing crowd segments, estimating crowd density and detection of unusual events are as follows:

Contribution 1: A new granular computing based dense crowd segmentation (GrCS) framework is proposed to infer crowd segments using the concept and principles of granular computing (GrC). GrC is incorporated in the framework to conceptualize crowd segmentation problem on different granularity similar to human cognition in problem solving, with the intention of mapping it into computationally tractable subproblems. Contrary to existing regular-grid representation (Fagette et al., 2014; Arandjelovic, 2008), the proposed GrCS framework studies the correlation among granules to represent structurally similar regions in crowd scene images to infer the crowd and background regions. This is essential because structures of background in the scene image can resemble crowd regions, which lead to vague outline between crowd and background. GrCS is scene-independent, and can be applied to dense crowd scenes with different physical layout.

Extensive experiments have been conducted on hundreds of real and synthetic crowd scenes. The results demonstrate that by exploiting the correlation among granules, one can outline the natural boundaries of structurally similar crowd and background regions necessary for dense crowd segmentation. To the best of my knowledge, this is one of the earliest works that uses GrC for dense crowd segmentation.

Contribution 2: The GrCS algorithm is extended to allow estimation of crowd density without tracking of features or segregation of individuals. As opposed to existing methods (Idrees et al., 2013; Marana, Velastin, Costa, & Lotufo, 1998), the proposed crowd density estimation approach partitions crowd scene images into irregular size granules

conforming to the boundaries of crowd and non-crowd regions. The underlying spatial information of each granule are exploited in a holistic manner to establish a direct mapping to the actual people counts. This caters for arbitrary distribution of crowd in different scenes (i.e. scene-invariant). Experimental results on standard public dataset demonstrate the effectiveness of using structurally meaningful granules for dense crowd density estimation.

Contribution 3: A novel framework to localize salient regions in crowd scene by transforming low-level motion features into global similarity structure is proposed. The structure allows the discovery of the intrinsic manifold of the motion dynamics in crowded scenes, which could not be captured by the low-level representation as to (Ali & Shah, 2007; Loy et al., 2012). Moreover, analysing the motion dynamic using global similarity to infer saliency in crowded scenes alleviate the need of (1) tracking, as the proposed approach exploits optical flow representation, and (2) prior information or model learning to identify interesting/salient regions in the crowded scenes.

Experimental results on public datasets demonstrate the effectiveness of exploiting global similarity structure to identify salient regions in various crowd scenarios that exhibit crowding, local irregular motion, and unique motion areas such as sources and sinks.

1.4 Organization of the Thesis

This chapter provides an overview of the work presented in the thesis. The remainder of the thesis is organized as follows:

Chapter 2 presents a review on existing literature that focuses on the strategies and approaches relevant to the three analyses that this thesis is focusing on, while discussing the main challenges and providing additional motivations for the proposed frameworks of

this thesis.

Chapter 3 presents the GrCS framework for crowd segmentation. It shows that exploiting the correlation among image granules at different levels of granularity are not only useful in outlining natural boundaries between crowd and background (i.e. non-crowd) regions, but also important as a meaningful primitive region to facilitate more robust and accurate crowd segmentation.

Chapter 4 explains the mechanism of granularity-based approach for crowd density estimation using contextual (i.e. spatial) information. Experiments are carried out to evaluate the effectiveness of the proposed approach in adapting to different public crowd scenes to estimate number of individuals in extremely dense crowds.

Chapter 5 provides detailed explanations on the proposed framework to identify and localize salient regions in a crowd scene. In particular, the chapter describes the transformation of low-level features extracted from crowd motion field into a global similarity structure. Experiments are conducted to demonstrate the effectiveness of the proposed framework in discovering the intrinsic manifold of the motion dynamic to identifying salient regions in various crowd scenarios.

Chapter 6 draws the previous chapters to a conclusion and recommends a number of areas to be pursued as future work.

CHAPTER 2: LITERATURE REVIEW

Visual crowd surveillance at large public events such as concerts, parades and rallies are common in cities worldwide. The mere existence of crowd has the prospect of progressing into a hazardous scene, for instance, the recent stampede in the Shanghai 2014 New Year's Eve revelry which claimed 36 innocent lives.[1] Alarmingly, with rapid urbanization around the world, the formation of crowd by chance is becoming a norm, e.g. crowds in train stations during rush hour. Along with the high frequency of crowd disasters and the growth of visual surveillance system at key crowd locations (e.g. train stations, markets and airports), crowd analysis in computer vision has recently play a growing role in visual surveillance.

Substantial effort has been spent driven by the practical demand, and it is becoming an important research direction (Junior et al., 2010). Given the broad and growing nature of crowd analysis in computer vision, this chapter narrows down the research scope by reviewing studies that address the major tasks associated with the analysis of dense crowd scenes. This chapter focuses on: (1) segmenting and localizing regions of dense crowd in a scene, (2) determining the density of people in a dense crowd scene and (3) crowd saliency detection. Specifically, the review in this chapter is structured into four subsections: strategies for dense crowd analysis (Section 2.1), dense crowd segmentation (Section 2.2), density estimation (Section 2.3) and saliency detection in dense crowd scenes (Section 2.4).

Some specific features and techniques used for visual analysis of crowd, such as optical flow and object tracking are not described thoroughly in this review. A literature review performed by Thida, Yong, Climent-Pérez, Eng, and Remagnino (2013) provides detailed studies on the aforementioned feature and techniques. Zhan et al. (2008) and

[1] BBC News: http://www.bbc.com/news/world-asia-china-30646918

Junior et al. (2010) provides comprehensive coverage on different strategies developed in computer vision techniques for crowd analysis. For crowd density estimation, comprehensive descriptions of the state-of-the-art approaches with emphasis on the methodologies and systematic evaluation can be found in (Loy et al., 2013). T. Li et al. (2015) highlight the techniques for crowded scenes analysis from 2010 onward.

2.1 Dense Crowd Analysis Strategies

Despite the practical significance of dense crowd analysis, the visual processes of dense crowd still pose tremendous challenges for computer vision. Particularly, the stochastic nature of individual in dense crowd tends to be highly challenging for traditional spatial-temporal representation (Chan, 2008). Computer vision algorithms have, for the most part, been restricted to visual analysis of sparse crowd scenes mainly due to the limitations of person detection and tracking. As density in the scene increases, the complexity increases and may become intractable. The complexities often manifest itself in partial or complete occlusion among individuals and complex events due to interactions among individuals in crowd, as discussed in Chapter 1 (Section 1.1). A significant degradation in the performance of analysis is usually observed in terms of detection and tracking, given that many existing methods rely on the ability to separate each individual from each other and from the background (Rodriguez et al., 2012).

Thus, visual analysis of dense crowd is a distinctive research problem that had emerged as an increasingly dedicated problem. Significant progress has been made in this field. The analyses of dense crowd scenes are commonly conducted at microscopic or macroscopic levels (Thida et al., 2013).

Microscopic Level Analysis: Inherently local, where it deals with the crowd as discrete individuals. The microscopic model depends on the analysis of motions of each

individual in crowd to achieve understanding of the whole crowd. This type of analysis generally commence by detecting the moving individuals present in the scene. Then, the detected individuals are tracked as they enter the scene till they exit the scene, and the tracked results are compiled for subsequent visual analysis of dense crowd (e.g. density estimation and anomaly detection). Such method works well on scenes that are relatively sparse (5-20 individuals) and is not appropriate to analyze dense crowd (Ali et al., 2013).

Macroscopic Level Analysis: A crowd is treated as a single entity. Macroscopic model is interested in the global motions of a crowd of individuals, without specifically analyze motions of each individual in a crowd. Holistic properties (e.g. instantaneous motions of the entire scenes) are usually utilized to learn the typical motion patterns in a crowd scene. This is the preferred approach to analyze both sparse and dense crowd (Thida et al., 2013).

In the following sections, the researches (i.e. dense crowd segmentation, density estimation and saliency detection) associated with the analysis of dense crowd scenes incorporating the aspects of microscopic and macroscopic analysis are discussed. Advantages and weakness of the many existing approaches are also highlighted.

2.2 Dense Crowd Segmentation

Dense crowd segmentation refers to the process of differentiating crowds from background regions (e.g. buildings, vehicles and trees). Generally, work in dense crowd segmentation assume that crowd is an agglomeration of pedestrians (B. Zhou et al., 2012). Even though each individual has their own goal destination and inclination, they appear to share common motion dynamics when observed over time in a crowded scene. This is due in part to the tendency of individuals to follow the dominant flow owing to the physical structure of the scene, and the social conventions of the crowd dynamics.

The complexity of dense crowd scenes often manifests itself in partial or complete occlusions among the individuals in crowd. The fact that each individual in dense crowd scenes is occluding each other blurs that boundary of crowd and non-crowd pixels in the scene. Therefore, dense crowd segmentation is commonly the basis for subsequent more complex task in analyzing dense crowd, such as crowd density estimation (Idrees et al., 2013) and crowd behavior analysis (Solmaz et al., 2012). This is also stated in (Idrees et al., 2015) and (Kang & Wang, 2014) that the localization of crowd segments is required prior to visual tasks such as behavior understanding for saliency detection. In some scenarios, crowd segmentation is applied prior to estimating the density of crowd (Idrees et al., 2013; Zhang & Li, 2012).

Recently, a significant amount of effort has been placed to develop models and strategies to localize dense crowd segments in public scenes using computer vision techniques. In this review, these models and strategies are divided into two categories: motion flow based model and feature based model.

2.2.1 Motion Flow Based Model

Crowd is generally studied with emphasis given on the evolution of its motions in an environment. Existing work on dense crowd analysis tends to exploit the collective coordination of crowd by analyzing crowd through analogies with studies in fluid dynamic (Moore et al., 2011; Ali & Shah, 2007; Shah, 2010) or treating a crowd as a collective entity (B. Zhou et al., 2012; Ali & Shah, 2008; Mehran, Oyama, & Shah, 2009; Hou & Pang, 2013). The main focus is to group regions with similar motion dynamics or coherency (Wu & San Wong, 2012; Rodriguez et al., 2012), such as illustrated in Figure 2.1. A number of approaches have been proposed for crowd segmentation. These studies lean towards analyzing dynamic crowd segments for crowd flow segmentation (Ali & Shah, 2007; Wu et al., 2009), crowd behavior understanding (Solmaz et al., 2012),

(a) Ali and Shah (2007) **(b)** Wu et al. (2009)

Figure 2.1: An illustration of two marathon sequences and the corresponding dense crowd segmentation results using motion flow based method proposed by (a) Ali and Shah (2007) and (b) Wu et al. (2009). The different colors in (a) represent different flow segments. Best viewed in color.

person tracking (Ali & Shah, 2008; Mazzon, Tahir, & Cavallaro, 2012), anomaly segmentation (Leach, Sparks, & Robertson, 2014) and crowd counting (Chan et al., 2008).

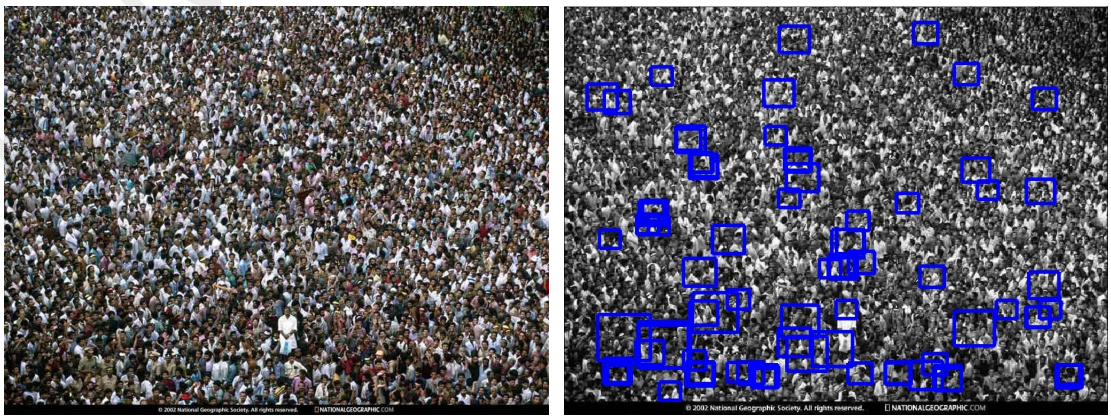
Most often, rather than computing the trajectories of individuals (microscopic), holistic approaches (macroscopic) represent crowd motion patterns using instantaneous motions of the entire scene such as the flow field (Ali & Shah, 2007, 2008; Mehran et al., 2010; Wu et al., 2009; M. Hu, Ali, & Shah, 2008). These flow fields are then combined with an agglomerative clustering algorithm (M. Hu et al., 2008) or Lagrangian particle dynamics (Ali & Shah, 2007, 2008) to partition crowd scenes into regions with similar coherent motion. There are, however some work which is based on tracking individuals and accumulating their trajectories over a period of time to obtain coherent motion (B. Zhou et al., 2012). Tracking approaches, regardless of whether they are using distance or model-based representations are very challenging in dense crowd scenes (Chongjing, Xu, Yi, & Yuncai, 2013). This is because the trajectories are highly fragmented with many missing observations due to the complex interactions, occlusions between individuals in the crowd and background clutters. Therefore, tracking in dense crowded scenes often incorporate scene or contextual information to enhance trajectory estimation (Dehghan, Idrees, Zamir, & Shah, 2014).

In another variation, some approaches perform background subtraction (Kong, Gray, & Tao, 2006; Dong, Parameswaran, Ramesh, & Zoghlami, 2007) to identify crowd seg-

ments. Such approaches are susceptible to false segmentation in cluttered environment with other moving entities (e.g. moving vehicles and waving trees), as well as limited to localizing crowd with variations in collective motion. Observations by Helbing, Molnar, Farkas, and Bolay (2001) highlighted that stationary crowd (e.g. spectators of a speech) implicitly influenced the motion flow of dynamic crowd, where crowd maneuver around stationary crowd to avoid collisions. Thus, it is of equal importance to include stationary crowd segments for a complete crowd surveillance system.

2.2.2 Feature Based Model

While the earlier discussed works are fixated on segmenting coherent motions as a cue of crowd on videos or image sequences, there is another branch of crowd segmentation research that exploits the holistic and collective description of crowd pattern, regardless of the motion variations. Due to severe inter-occlusions and perspective distortion in dense crowd scene, appearance-based approaches which include head and shoulder segmentation are still in their infancy stage. This is an ongoing research problem (Idrees et al., 2015). Figure 2.2 shows an example of a dense crowd scene where individuals in crowd are severely inter-occluded and mostly cannot be detected.



(a) Dense crowd scene. **(b)** Person (head) detection

Figure 2.2: Person (head) detection result using state-of-the-art method (Felzenszwalb et al., 2008). The blue bounding boxes signify the detections results. False positive and fail detections are evident in the image. Best viewed in color. ((Felzenszwalb et al., 2008))

To alleviate the need of person detection in a crowd, imagery of crowd scenes is partitioned into regular pixel-grid with the purpose of achieving local texture consistency and is treated as a texture analysis problem. A study by Marana et al. (1998) verifies that crowd regions carry strong cue of texture variations. Arandjelovic (2008) proposes an image-based crowd segmentation method using low-level local feature from single crowd image. Each pixel response is defined by using multi-scale pixel-grid, where the computation of the probability of a pixel-grid being a crowd region is based on a predefined average number of SIFT word segmentation per image area. Example results are as illustrated in Figure 2.3. Using similar approach, Idrees et al. (2013) partition crowd scene into pixel-grid to construct a confidence map of crowd regions. In another study, Fagette et al. (2014) perform crowd segmentation by retrieving multi-scale pixel-grid texture features from crowd scene. Binary classification is conducted to infer crowd regions in image. Since these methods use regular pixel-grid, the representation is not adaptive to the random distribution of crowd perimeters in real-world scene. Also, it is unclear how well they can be generalized to arbitrary crowd scenes. The number of layers in multi-scale pixel grid is scene-dependent; it has to be empirically defined for each public crowd scene to optimize adherence to the arbitrary crowd distribution.

Figure 2.3: Sample results of dense crowd segmentation where regions containing crowd are segmented using method as proposed by Arandjelovic (2008). The true positives are highlighted in green whereas the false positives are represented by the red areas. Best viewed in color. ((Arandjelovic, 2008))

2.2.3 Discussion

Existing feature based crowd segmentation model (Arandjelovic, 2008; Idrees et al., 2013; Ghidoni, Cielniak, & Menegatti, 2013; Fagette et al., 2014) infer crowd segments by learning the textures of crowd scenes either using regular pixel-grid or overlapping multi-scale pixel-grid (i.e. numerous range of neighboring pixels) at each pixel. This is to obviate the difficulties to segregate individuals in dense crowd scenes due to appearance variations of crowd and poor resolution. In spite of the promising results, the use of pixel-grid imposes some constraints on inferring crowd segments. In the former, crowd images are divided into regular pixel grids where an optimized boundary adherence of crowd segments across different scene is difficult to achieve. In the latter, an antecedent version of the regular pixel-grid, namely, multi-scale pixel grid is proposed to cope with crowd variation across different scenes. Since it is leveraging on its antecedent, conformation to varying crowd segments remains unresolved. With a smaller pixel-grid, localization accuracy is better with less probability of patch consisting both crowd and background regions; whereas a larger pixel-grid covers wider regions for analysis of structure (Arandjelovic, 2008; Kang & Wang, 2014).

The dense crowd segmentation method proposed in this thesis (see Chapter 3) exploit the correlations among image granules of varying sizes with the hope to alleviate the aforementioned constrains. Importantly, it simplifies each public crowd scene into structurally meaningful granules to optimize adherence to the arbitrary crowd distribution for dense crowd segmentation.

2.3 Crowd Density Estimation

Not all events with large gathering of people are conducted in an enclosed venue with turnstiles where crowd density estimation can be administered seamlessly. And for some events that are held in an open space area such as parades or political protest, employing

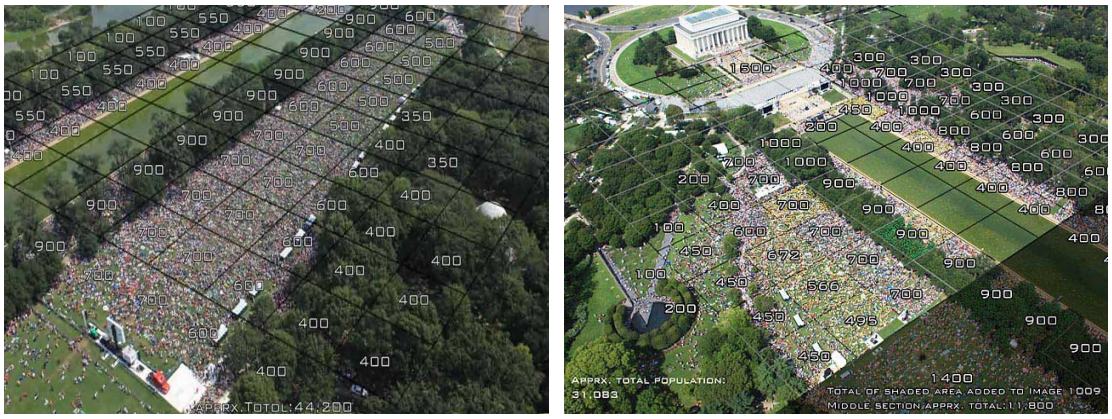


Figure 2.4: Crowd density estimation using Jacob's method. Grids are overlaid on the crowd scene to compute the average number of individuals per square meter, and multiplying with the total squares to determine the approximate number of individuals in a scene. Image source: Digital Design & Imaging Service Inc. (2015)

professionals to conduct human counting is infeasible. Nevertheless, estimating density of crowd is of utmost importance to better administer the well-being of crowd as a whole, development of public space design and accurate documentation of historical events. The Hillsborough disaster (Taylor, 1990) is an example of the consequences of overcrowding. Such tragedy could be avoided if a more effective crowd control system was enforced. Crowd density estimation system can be a highly beneficial tool to monitor the density of crowd to enable more effective crowd control.

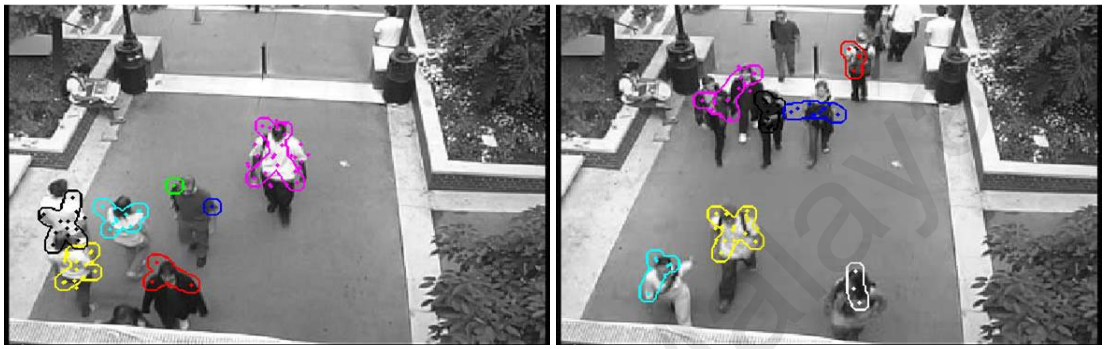
In 1967, Herbert Jacobs proposed to estimate crowd density by getting an average of individuals per square meter, and multiply that by the total squares as depicted in Figure 2.4. The approach has modernized and led to a paradigm shift in the way to estimating density in crowd (Weiss, 2013). Determining the density of individuals in crowded scenes has been investigated in numerous studies in computer vision. The aim of most of the studies that focus on this task is to deliver precise estimation of individual within a scene or a given spatio-temporal region of a scene. In this section, the studies are categorized into two subcategories: object-level and texture-level analysis.

2.3.1 Object Level Analysis

Existing work on crowd density estimation depends mainly on collective motion and appearance cues, with respect to the type of inputs (i.e. crowd video sequences or single crowd image). Different techniques are adopted to cope with crowd scene of varying density. The greater density of crowd in a scene, the more complicated the task to estimate crowd density where dynamic occlusions come into picture. It is infeasible to discern different person and ones' body parts when a person may only be occupying few pixels (Idrees et al., 2013) and further rendered by background clutter. Nevertheless, a significant amount of density estimation algorithms infer person count from local object detector. For instance, framework that performs clustering of coherent trajectories to represent a moving entity, and inferring number of individual in the scene by Rabaud and Belongie (2006). This approach is limited to crowd scenes with sparse crowd where continuous sets of image frames are accessible. The results presented in their work have shown promising performance in Figure 2.5a when individuals are disconnected from each other. However when individuals in crowd scenes are closely positioned with each other, trajectories are incorrectly merged such as depicted in Figure 2.5b. This is due to the phenomenon of collective motion occurring between moving interacting entities.

Using an analogous perception, M. Li, Zhang, Huang, and Tan (2008) estimate the numbers of people in crowd by implementing foreground segmentation and head-shoulder detection approach. The proposed method was intended to address stationary crowd, where subtle motions of individual is crucial and deeply relied on in defining foreground segments. Nonetheless, the proposed framework is susceptible to inter-occlusion between individuals, particularly prominent in a dense crowd scene. Ge and Collins in (Ge & Collins, 2009), proposed a Bayesian marked point process to detect individuals in crowd where clear silhouette of individuals is required for accurate projection to a trained

(a) Accurate clustering of the trajectories of independent individuals.



(b) Inaccurate clustering of the trajectories of independent individuals.

Figure 2.5: Sample results of density estimation on sparse crowd scene where coherent trajactories are agglomeratively clustered to deduce the number of persons. The clustered trajectories are denoted with different colors (i.e. black, blue, green, red, white, yellow, cyan and pink). (a) Trajectories of independent individuals are accurately clustered where the number of resulting clusters denotes the density of individuals. (b) Inter-occlusions between individuals lead to inaccurate merging of the trajectories of multiple individuals (left: black cluster, right: pink cluster). Best viewed in color. ((Rabaud & Belongie, 2006))

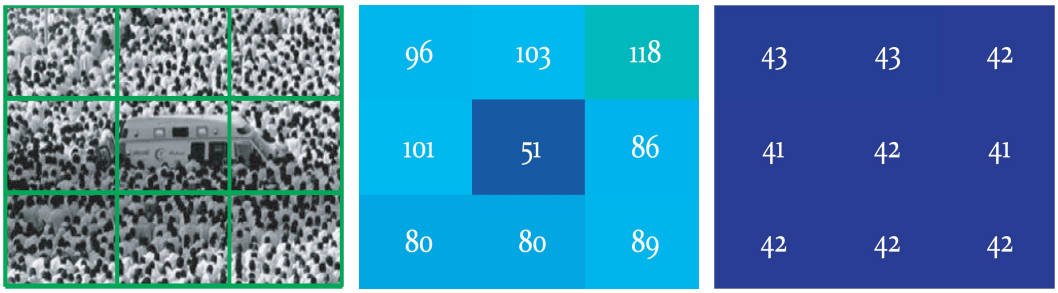
set for accurate detection and counting of individuals. In another study, Ge and Collins (2010) uses a generative sampling-based approach that leverage on multi-view geometry to achieve density estimation of individuals in crowd. The work assumes that individuals in a crowd retain a certain space with each other. Thus, individuals in the scene should not be occluded from all viewing angle. This approach tends to generate accurate density estimation only within the bounds of the previously mentioned assumption.

2.3.2 Texture Level Analysis

Alleviating the need to detect each person in a crowd, some works (Marana et al., 1998; Davies et al., 1995; Idrees et al., 2013; K. Chen et al., 2012; Schofield, Mehta, & Stonham, 1996; Tan, Zhang, & Wang, 2011; Liang, Zhu, & Wang, 2014) uses low level crowd

features (appearance cue) formed based on the collectives of crowd to estimate crowd density. Marana et al. (1998) presented a method based on texture analysis to estimate crowd density, where the estimation is given in terms of discrete ranges (i.e. very low, low, moderate, high and very high). Their objective was to challenge scenes of dense crowd where each individual is greatly occluded. They assumed that crowd scene of high density tend to illustrate fine textures, whereas crowd scene of low density are mostly made up of coarse patterns.

Crowd density estimation by Davies et al. (1995) is one of the earliest works that uses regression approach to learn the relationship between global features (e.g. number of edge pixels) and density of individuals. Similarly, works by Chan et al. (2008) as well as Chan and Vasconcelos (2012) propose to extract dynamic texture from homogeneous motion crowd segments and focus on learning mapping between large set of feature responses and density. A problem commonly encountered in regression based density estimation is perspective distortion, where individuals who are closer to the camera view appear larger than those who are positioned further away from the camera (Loy et al., 2013). The problem is exacerbated when single regression function is used for the whole image space. To address this problem, perspective normalization plays a key role by bringing the perceived size of individuals at different depths to the same scale. Another approach is to divide the image space into different cells and each cell is modeled by a regression function to mitigate the influence of perspective distortion. K. Chen et al. (2012) proposed a multi-output regression approach to estimate crowd density in sparse crowd images. Low-level features extracted are shared among spatially localized regions to achieve more accurate counts prediction, indicating correlation between local regions of crowd scene is crucial. Idrees et al. (2013) estimate the number of individuals given single dense crowd image by leveraging the harmonic textures elements of crowd from finer scales and appearances cues to approximate the density of crowd per image



(a) Pixel grid patches. **(b)** Ground truth. **(c)** Density estimation using (Idrees et al., 2013)

Figure 2.6: Crowd image partitioned into nine pixel grid patches (outlined in green) for regression based density estimation using method as proposed by Idrees et al. (2013). Density of individuals in patches with crowd and background (i.e. vehicle) and patches consisting of crowd only are inaccurately estimated. Best viewed in color. ((Idrees et al., 2013))

patch. The system uses regression approach to infer the count of individuals per patch and multi-scale random fields to refine the counts of individuals per image. Despite the promising results, this approach is constrained by the pixel grid patches such as depicted in Figure 2.6. It is observed that density of individuals in the patches with both crowd and vehicle are inaccurately estimated to have comparable number of individuals with patches consisting crowd only.

In another variations, Lempitsky and Zisserman (2010) model the density function over pixel grids, where integral over any region in the image would yield the density of object within. Kong et al. (2006) uses feed-forward neural network to map the correlation between feature histogram from low-level features and number of pedestrian.

2.3.3 Discussion

Over the years, density estimation has traditionally been focused on scenes containing low density of people. However, the interest in the areas of dense crowd density estimation has increased recently in the computer vision community.

Most of the aforementioned density estimation approaches have been constructed primarily to deduce density of sparse crowd scenes. Individuals in dense crowd, generally, do not uniformly distribute over a scene, but clump together as clusters or groups. Hence,

the approaches are subjected to the limitation of person detection and tracking (object level analysis), as well as pixel grid when coping with perspective distortion in texture level analysis.

To address the aforementioned constrains, Chapter 4 presents a novel framework to determine density of individuals by exploiting irregular patches in dense crowd scenes. These patches adhere to the outline of crowd and background regions (Chapter 3). A set of discriminative spatial information of each patches are extracted to estimate density of dense crowd scenes.

2.4 Crowd Saliency Detection

The formation of crowd and mass gathering often poses challenges to public safety if it is not handled effectively, particularly when panic arises among surging individuals (Helbing, Farkas, Molnar, & Vicsek, 2002). Therefore, amongst the major goal of computer vision systems is to detect and analyze the motion dynamics of crowded scenes, in the hope towards profiling and identifying salient motion behaviors which could lead to potential unfavorable events.

Existing crowd saliency detection methods can be divided into two major approaches. The first approach analyzes crowd behaviors or activities based on the motion of individuals, where tracking of their trajectories is required (Makris & Ellis, 2005; X. Wang, Tieu, & Grimson, 2006; Rodriguez, Ali, & Kanade, 2009; Rodriguez, Sivic, et al., 2011; Nedrich & Davis, 2010; B. Zhou et al., 2012). Another approach characterize crowd scenes as a collection of local motion estimates instead of a collection of object, i.e. holistic approach (Ali et al., 2013).

2.4.1 Object-centric Approach

Commonly, the object-centric approach keeps track of each individual motion and further applies a statistical model of the trajectories to identify the semantics or geometric

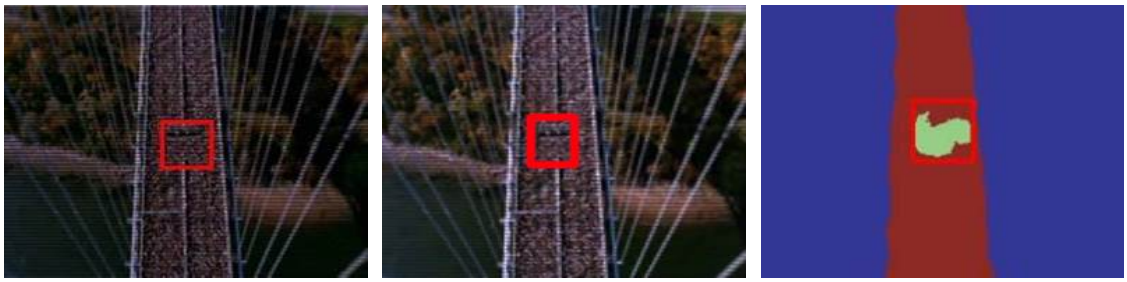
structures of the scene, such as the walking paths, sources and sinks. Then, the learned semantics are compared to the query trajectories to detect anomaly. These methods works well and produce promising results for sparse crowd scenes (i.e. with approximately 5-20 individuals) (Ali et al., 2013).

Without using a statistical measure of typicality, Dee and Hogg (2004) use the understanding of the way individuals navigate to identify individual that deviate from the goal-directed behavior. X. Wang et al. (2006) propose an unsupervised learning framework to learn semantic scene models using the tracking information. Abnormalities in the scenes are detected using the learned semantic scene models. W. Hu et al. (2006) learn the motion patterns in a scene by robustly track multiple objects, with the aim to detect anomaly and predict behavior. Similarly, B. Zhou et al. (2012) learn the collective behavior pattern of individuals in crowd scenes given their trajectories to infer their past behavior and predict the future behaviors.

While in principle individuals should be tracked from the time they enter a scene, till the time they exit the scene to infer such semantics, it is inevitable that tracking tends to fail due to occlusion, clutter background and irregular motion in crowded scenes. It may even become intractable with moderately dense crowd (Ali et al., 2013). The complexity of object-centric approaches increases disproportionately depending on the density of individuals in crowd scenes. The performance of saliency detection tends to deteriorate in dense crowd scenes, where target tracking is extremely challenging. Therefore, the aforementioned methods work well, up to a certain extent, even in sparse crowd scenes.

2.4.2 Holistic Approach

Another crowd saliency detection paradigm is based on the motion dynamics of crowd scenes (Ali & Shah, 2007; D.-Y. Chen & Huang, 2011; Loy et al., 2012; Solmaz et al., 2012; Zhu, Liu, Wang, Li, & Lu, 2014). This class of approaches obviates the challenges



(a) Crowd in a marathon scene. **(b)** Loy et al. (2012) **(c)** Ali and Shah (2007)

Figure 2.7: Sample results of saliency detection in dense crowd scene using method proposed by Loy et al. (2012) and (Ali & Shah, 2007). (a) Marathon sequence, where the abnormal region (enclosed in the red bounding box) is simulated by inserting synthetic instability into the original video. (b) Salient region detected by exploiting the instability information as proposed by Loy et al. (2012). (c) Salient region detected using the global motion saliency detection method based on spectral analysis as proposed by Ali and Shah (2007). Best viewed in color.

of detecting individuals and instead focuses on learning crowd motion models that capture the variations in local spatio-temporal motion patterns of crowd scenes. Finding interesting regions in a given scene is generally accomplished by firstly learn an activity model of the scene, followed by using the learned model to identify the anomalies (Kuettel, Breitenstein, Van Gool, & Ferrari, 2010; Hospedales, Li, Gong, & Xiang, 2011; B. Zhou et al., 2012; Rodriguez, Sivic, et al., 2011). In another variation, the flow field is clustered to detect typical motions in crowded scenes (M. Hu et al., 2008), or use a hidden Markov model to learn the inherent dynamics of the motion patterns for detection of saliency in crowds (Andrade, Blunsden, & Fisher, 2006). Ali and Shah (2007) apply the Lagrangian particle dynamics based on the crowd flow field to estimate the stability of a particular region. Their method able to detect regions with unstable motion by discovering the abnormality in the segmented flow fields (example result as shown in Figure 2.7c). Since the aforementioned methods use only the direction and speed as the motion features, their scenarios are limited to abnormal events that are varied in terms of motion direction and speed.

Detection and localization of salient regions by using spectral analysis is proposed by Loy et al. (2012). In contrast to other methods, their method suppress dominant flows with

a focus on the motion flows that deviate from the norm (example result as shown in Figure 2.7b). Solmaz et al. (2012) propose a linear approximation of the dynamical system to categorize different crowd behaviors using the eigenvalues over an interval of time. Their methods show promising results in detecting and classifying five different scenarios of saliency, which includes the bottleneck, lane, arch, fountainhead and blocking.

There are also other approaches that adopt learning methods to interpret crowd dynamics for saliency detection. Kratz and Nishino (2009), for example, propose to model a 3D Gaussian distributions representation of spatio-temporal motion patterns. This is then fed into a variant of Hidden Markov Model to discover the relationships between these patterns. Saliency is defined as statistical deviations within the video sequences of the same scene. In the more recent works by Mahadevan et al. (2010) and W. x. Li et al. (2014), a joint models of appearance and dynamics is proposed, known as the dynamic textures (DT). Hierarchical mixtures of DT models are then performed, where the spatial and temporal saliency scores are integrated across time, space and scale with a conditional random field (CRF). Here, saliency is defined as events of low probability with respect to a model or normal crowd behavior.

One of the foremost challenges in crowd saliency detection is the need of large amount of data to enable good learning for discriminative saliency detection. In (Ihaddadene & Djeraba, 2008), a non-learning method for crowd dynamic analysis is proposed to mitigate the need of requiring a huge amount of data for accurate learning. Their proposed method detects saliency by observing the deviations of features between a set of points-of-interest (POI) over a time series. Although the proposed non-learning method provides convenient solution, it is restricted to a particular behavior or event such as detecting collapse flow near escalator exits and may not be ideal in dealing with the complexity of real-world scenarios.

2.4.3 Discussion

Generally, large amount of data is required to enable good supervised / unsupervised learning for discriminative or generative crowd models. However, a major challenge in the context of crowd analysis in surveillance applications is the lack of abnormal or ground truth events for training. The typical and normal individual behaviors in crowd scenes are often known a-priori, whereas abnormal activities in crowd are erratic (Loy, 2010). Even if abnormalities of a crowd scene can be comprehensively inferred, the learned model is scene-dependent and not adaptive to different public crowd scenes. To address this problem, Chapter 5 will present an approach that transform low-level motion features into global similarity structure to uncover the intrinsic manifold of the motion dynamics. The extrema in the intrinsic manifold serve as the indicator of saliency. It is therefore requires no tracking or model learning to identify salient regions in dense crowd scenes.

2.5 Summary

The preceding reviews and discussions have covered essential studies in the literature regarding visual crowd analysis. Specifically, various state of the art approaches for crowd segmentation, density estimation and saliency detection have been reviewed. This chapter has also discussed several open problems and limitations that need to be solved when dealing with dense crowd scene. Firstly, most conventional computer vision algorithms are object-centric, where detecting and learning the motion of moving individuals in a scene is important. It serves as motion priors that can be used to enhance subsequent tasks for visual analysis such as density estimation and saliency detection. This method tends to fail given individuals in dense crowd scenes are likely to be densely packed together. Secondly, the inherent constraints of pixel grid patches has never been attended to date. The notion of simplifying scenes into meaningful atomic regions by exploiting the

correlations among image features is generally unprecedented for dense crowd segments and to determine the density of individuals. Thirdly, saliency detection in dense crowd scenarios mostly uses low-level motion features that may be prone to false detection as a result of ambiguity in feature space. In subsequent chapters of this thesis, algorithms are formulated to address these constrains.

CHAPTER 3: GRANULAR COMPUTING BASED DENSE CROWD SEGMENTATION (GRCS)

Dense crowd segmentation is important in serving as the basis for a wide range of crowd analysis tasks such as density estimation and behavior understanding. However, due to inter-occlusions, perspective distortion, clutter background and random crowd distribution, localizing dense crowd segments is technically a very challenging task (discussed in Section 1.2).

To this end, this chapter proposes a novel granular computing (GrC) based approach for dense crowd segmentation. The aim is to simplify dense crowd scenes to alleviate the difficulty of defining the natural boundaries between crowd and background (i.e. non-crowd) regions. Unlike existing crowd analysis approaches (Fagette et al., 2014; Arandjelovic, 2008), the problem of dense crowd segmentation is decomposed into a family of sub-problems, denoted by granules in the proposed method. Granules are constructed by finer granules based on similarity and distinguishability (Zadeh, 1996). Specifically, by exploiting the correlation among pixel granules, the structurally similar pixels are able to be aggregated into meaningful atomic structure granules. This is useful in outlining natural boundaries between crowd and background (i.e. non-crowd) regions. From the structure granules, the granular computing based dense crowd segmentation (GrCS) infer the crowd and background regions by granular information classification. In contrast to existing methods (Arandjelovic, 2008; Fagette et al., 2014), GrCS is scene-independent, and can be applied effectively to crowd scenes with a variety of physical layouts and crowdedness.

The rest of the chapter is organized as follows: Section 3.1 introduces the intuition and motivation behind the proposed dense crowd segmentation approach. Section 3.2 describes the proposed framework of dense crowd segmentation by modeling crowd scenes

with granular computing (GrC). The experimental results are presented and discussed in Section 3.3. Specifically, the effectiveness of the proposed framework in dense crowd segmentation is evaluated using hundreds of real and synthetic dense crowd scenes. This is followed by the possible future work and conclusion in Section 3.4.

3.1 Dense Crowd Segmentation

In this chapter, the correlation among image granules at different levels of granularity is exploited with the hope that granulation can alleviate the constrains of pixel-grid approach (Arandjelovic, 2008; Idrees et al., 2013; Fagette et al., 2014) as discussed in Section 2.2.3.

The dichotomy articulated by Moravec (Moravec, 1988) between humans and machines regarding the easiness and complexity in solving different problems remains valid today. Specifically, machines perform poorly in tasks that are seemingly effortless and natural for humans (i.e. recognizing crowd regions), but can easily solve problems that humans find challenging (i.e. numerical computation). One key advantage of the human mind has over a machine in cognition is the ability to segment visual information into meaningful units of analysis effortlessly (Hendee & Wells, 1997). More remarkably, this is achieved in vivid detail; disregarding the orientation, color intensity and deformation present. This structured problem solving ability of human cognition is transferred into dense crowd segmentation system in this chapter, with the aim of alleviating the complexity to infer crowd segments.

Interestingly, granular computing (GrC), an emerging computing paradigm of information processing (Pedrycz, 2001), simulates human cognitive process by enabling abstraction on the essential details at different granularities. That is, correlations among granules are explored to solve various research problems in computer. For instance, Pal, Uma Shankar, and Mitra (2005) apply granular computing (GrC) together with rough sets to perform grayscale image segmentation. Their method defines non-overlapping

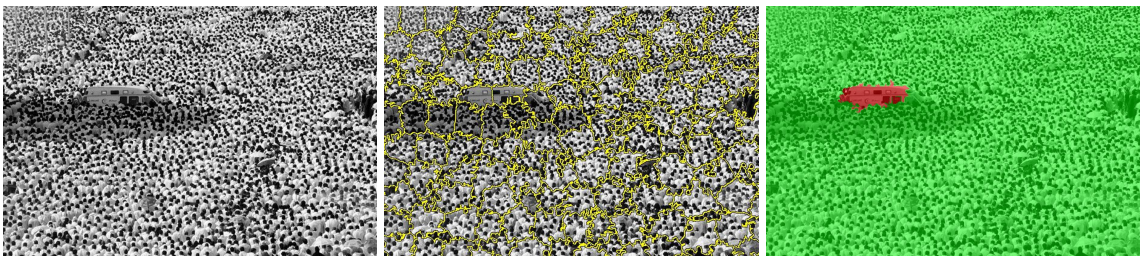


Figure 3.1: GrCS: Granular Computing based Dense Crowd Segmentation. (Left) dense crowd scene image. (Middle) image segmented into structurally-similar atomic clusters (structure granules), shown as regions within yellow outline. Perimeters of crowd and background are distinctively separated. (Right) crowd and background regions segmentation achieved via classification of structure granules. A vehicle is outlined and classified as background region (shown as red overlay). Best viewed in color.

pixel-grid of different sizes as granules to quantify the object-background regions in images. Rizzi and Del Vescovo (2006) propose to decompose each image into segments (i.e. granules) and map the correlation among image segments for image classification. The method performs abstractions to cope with a wide set of problem instances of image classification. The underlying idea of GrC is the use of classes, groups or clusters of elements denoted as granules (Y. Yao, 2000). These granules are drawn together by similarity and distinguishability (Zadeh, 1996). So unlike conventional approaches (Fagette et al., 2014; Arandjelovic, 2008), the concept of GrC is incorporated in the proposed approach in the form of granules, thereby, honoring the correlations of structures in dense crowd scenes from pixel level to crowd and background level. This is to mitigate the effects of issues, such as context variations of crowd, cluttered background and unconstrained physical layout of the environment, for an effective dense crowd segmentation. The utilization of granules obviates the difficulty to segregate individuals in dense crowd due to context variations of crowd by enabling inference of crowd and background regions based on local structures. To circumvent the effects of cluttered background and unconstrained physical layout of the environment, it is believed that the key is to study the correlations among granules to represent structurafdlly similar regions in crowd scene images.

The notion of simplifying an image scene into structurally meaningful atomic re-

gions (i.e. granules) is generally unprecedented in the existing crowd segmentation studies. It is important to have granulation that is able to adapt in different crowd structures in scenes due to varying crowdedness, perspective distortion, severe inter-occlusion and cluttered background for a better dense crowd segmentation. As an example, Fig. 3.1 illustrates a dense crowd scene with severe inter-occlusion between individuals and the scale of individuals vary drastically due to the perspective and position of camera. Even so humans are able to distinguish the vehicle within the crowd with ease. Similarly, by using the proposed method, the objective is to have granules (i.e. regions within the yellow outline) that encompass only a single context (i.e. crowd or background), as shown in Fig. 3.1 (Middle). This will serves as a meaningful primitive region to infer the corresponding context (as shown in Fig. 3.1 (Right)). Accordingly, the vehicle (red overlay) surrounded by a swarm of crowd (green overlay) can be effectively singled out despite severe occlusion and highly textured scene.

3.2 Proposed Dense Crowd Segmentation Framework

The key steps of granular computing based crowd segmentation (GrCS) framework are illustrated in Figure 3.2, where granules are the basic elements. Each level represents different levels of granularity, i.e. pixel, structure and foreground / background granules, which will be detailed in the subsequent sections. This is to simulate the ability of humans to conceptualize at different granularity levels with the intention of mapping problems into computationally tractable subproblems.

In this context, a dense crowd image, $I = [\boldsymbol{\upsilon}_{ps}] \in \mathbb{R}^{N \times S}$, where N is the number of pixels in an image and S is the number of features for each pixel, p. Each pixel, p, in an image is the basic granule (i.e. pixel granule), represented as a feature vector, $\boldsymbol{\upsilon}_{ps} = \left(v_{p1}, \ldots, v_{ps}, \ldots, v_{pS}\right)^{\top} \in \mathbb{R}^{N \times S}$, where $p = \{1, \ldots, N\}$ and $s = \{1, \ldots, S\}$. The feature vector, $\boldsymbol{\upsilon}_{ps}$ is formed by the concatenation of S features. Aggregation of the pixel

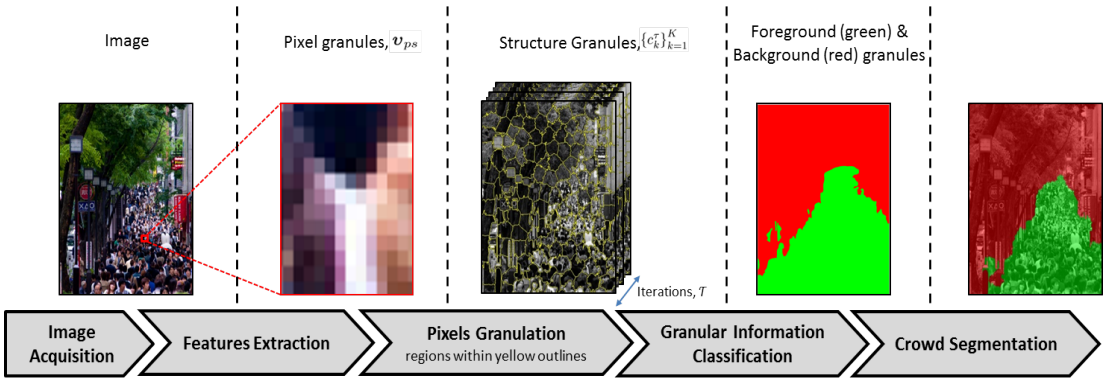


Figure 3.2: GrCS: granular computing based dense crowd segmentation framework. An illustration of the key steps and the different levels of granularity of image in granular computing based dense crowd segmentation (GrCS). Best viewed in color.

granules (granulation process) with similarity of feature vector, $\boldsymbol{\upsilon}_{ps}$, will form a higher level set of granules (i.e. structure granules). These structure granules are anticipated to be structurally coherent atomic regions in the image that conform to the natural boundaries between different structures of crowd and background. The key idea of the atomic regions is to have a pixel aggregation process versatile to different crowd scenes, and so this will best categorize the diverse structures in the scene for robust dense crowd segmentation. From the structure granules, the dense crowd segmentation task is posed as a classification problem to construct granulated view of foreground (i.e. crowd in the context of this chapter) and background (e.g. sky, buildings, grasses etc.) granules.

3.2.1 Pixel Granules

The finest level of granules represents the most basic aspect of dense crowd scenes, which is the pixel information: pixel intensity and spatial position in the image plane. However, due to the complexity of discerning cluttered background from crowd, texture features are introduced in this proposed framework to increase the discriminative ability for texture differentiation. This is because background region such as building, can be easily misinterpreted as crowd region (as shown in Figure 3.3). Co-occurrence of multiple features, v_{ps}, is thus essential to complement the insufficiencies of other features. Similar strategy is used by humans where one's cognition uses existing information to understand a new

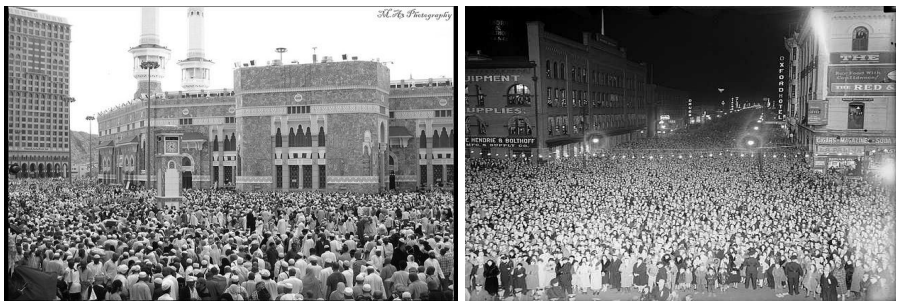


Figure 3.3: Example of dense crowd images where cluttered background regions (e.g. buildings) can blur the boundary between crowd and background regions. These cluttered background regions can be easily misinterpreted as crowd region as well.

subject matter.

In this proposed approach, the texture features are represented by the widely used *Local Binary Pattern (LBP)* (Ojala et al., 1996) and *Local Range of Intensity (LRI)*. Nevertheless, the proposed framework is not restricted to these sets of features employed in this chapter. Diverse sets of features can be exploited to enhance and adapt to various image segmentation task.

3.2.1.1 Local Binary Pattern (LBP)

LBP is computationally simple yet a practical grey-level invariant approach to summarize local grey-level structure. LBP is adopted to capture the microstructure of local region by which the raw low-level spatial pattern of dense crowd is analyzed. Employing LBP to capture the dense microstructures in crowd regions, such as lines and edges formed by a mass of crowd can serve as a good indicator of the presence of crowd. However, real world microstructures can occur at arbitrary orientations due to varying illumination conditions (Ojala, Pietikäinen, & Mäenpää, 2002). In this proposed approach, an extended version of LBP operator known as uniform patterns (Ojala et al., 2002) is thus implemented to cope with variance in rotation of captured microstructures.

Given pixels within a dense crowd image, I, a 3×3 circularly symmetric local neighborhood, i.e. 8 sampling points centering each pixel of interest is used (as illustrated in Figure 3.4). The sampling points are subtracted against the value of the corresponding

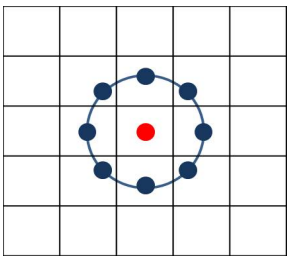


Figure 3.4: Example of the 3×3 circular neighborhood used to calculate a Local Binary Pattern (LBP). Red dot: pixel of interest. Blue dot: sampling point. Best viewed in color.

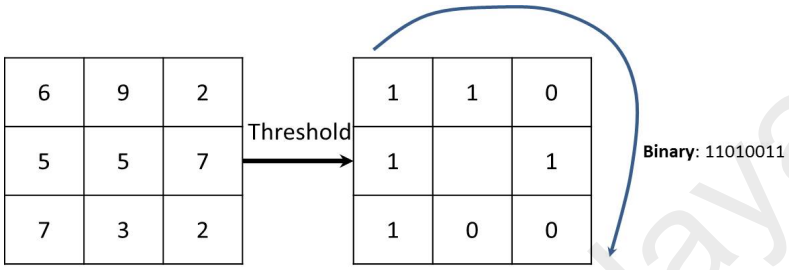


Figure 3.5: Example of the LBP operator by Ojala et al. (1996). (Left) A 3×3 circular neighborhood where the values indicates pixel intensities. (Right) The 8 sampling points centering the pixel of interest are threshold against the value of the corresponding pixel of interest. The resulting positive values are encoded with 1, and 0 otherwise. The binary values associated with the local neighborhood are concatenated in a clockwise direction (blue arrow) to form a binary pattern. Best viewed in color.

pixel of interest, where the resulting positive values are encoded with 1, and 0 otherwise. The corresponding binary values associated with the local neighborhood is concatenated in a clockwise direction (starts from its top-left neighbor) to form a binary pattern (as shown in Figure 3.5). A binary pattern is called uniform if it contains at most two $1-0$ or $0-1$ transition. For example, the binary pattern 00001000 is uniform whereas 11001101 is not. For uniform pattern LBP, there is a separate bin for each uniform pattern and all non-uniform patterns are assigned to a single bin. Texture descriptors of uniform pattern LBP correspond to the histogram formed by uniform and non-uniform binary pattern bins.

3.2.1.2 Local Range of Intensity (LRI)

The Local Range of Intensity (LRI) is defined as the difference between the extrema (maximum and minimum) intensity values of a local neighborhood centering each pixel of interest. The notion of using local intensity variation to solve visual analysis problem

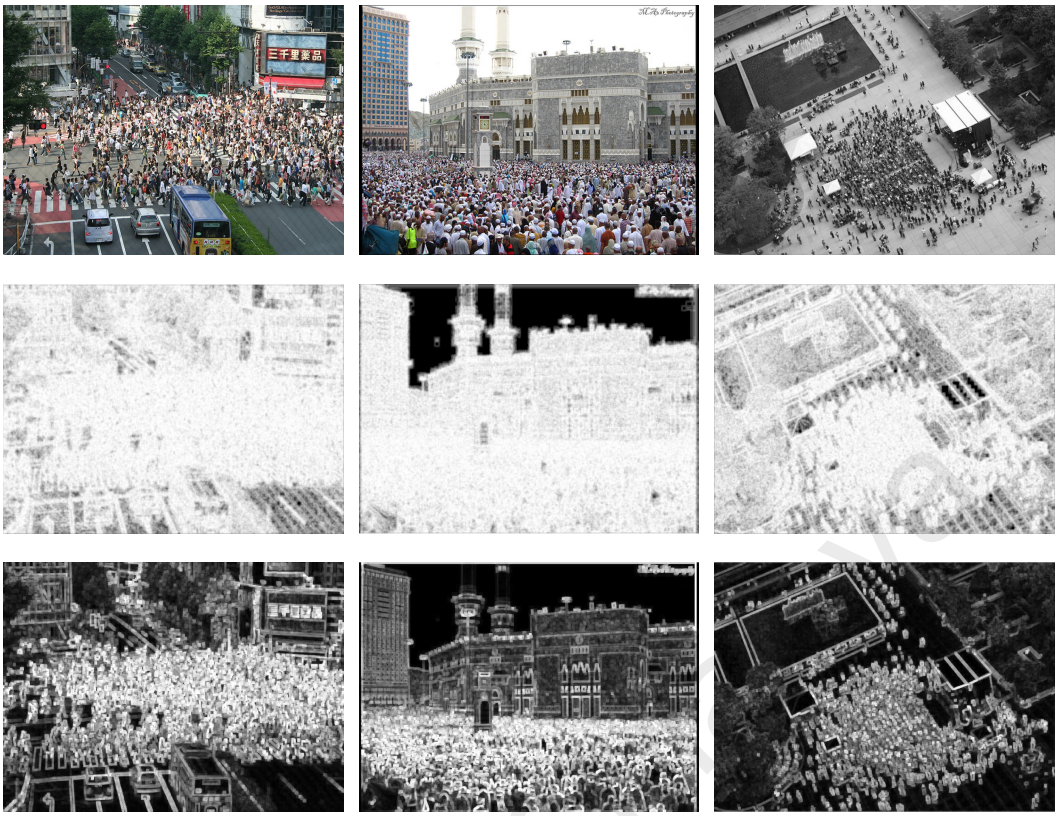


Figure 3.6: (Top row) Example crowd scene images. (Middle row) Entropy images using 5×5 neighbourhood. (Bottom row) Images of local range of intensity (LRI) using 5×5 neighbourhood. Best viewed in color.

in computer vision has been used by several researchers, such as J. Chen et al. (2008) and B. Wang, Li, Yang, and Liao (2011) for face detection and texture analysis.

As illustrated in Figure 3.6 (Top row), crowd segments tend to exhibit larger range of intensity variation in comparison to background (i.e. non-crowd) regions, mainly due to varying individual appearances. Instead of using the conventional entropy measure (Shannon, 2001), LRI is deemed more effective in quantifying the information content (statistical randomness) of local regions based on intensity variation in crowd scenes. Figure 3.6 (Middle row) shows that this is because conventional entropy is susceptible to image noise and background clutters such as grass, trees and buildings which produce similar entropy variation. However, as demonstrated in Figure 3.6, by adopting the LRI feature, the hurdle of discriminating crowd regions from textured background in existing literatures can be relaxed.

Figure 3.7: Sample background structure granules with variabilities in terms of illumination and texture patterns. Best viewed in color.

3.2.2 Structure Granulars

Crowdedness and the distribution of crowd in crowd scenes are rarely uniform due to the different physical layout of the environment (e.g. cinema, stadium and train station) and / or the viewpoint of the scene captured. Worse still, the textures of background (e.g. building structures and trees) and crowd (as a result of gait, clothing and shape of person) vary drastically, as illustrated in Figure 3.7 and Figure 3.8, respectively. It, thus, can lead to vague boundaries between crowd and background (as shown in Figure 3.6 (Top row)). On a finer scale, the variability of crowd region tends to corresponds to a unison structure (Idrees et al., 2013) as shown in Figure 3.8. The structures can be intimately governed by the structure granules to outline the perimeters of coherent crowd structure and background.

To this end, the correlations among pixel granules are explored for granulation. The aim is to form structurally uniform structure granules adhering to the natural edges of crowd scenes for analysis. This is analogous to how human brains perceive and process visual information; one does not focus on individual pixels, instead, grouping them into semantically meaningful forms to understand the image. In GrC, granulation process is the aggregation of smaller and lower level granules into a larger and higher level granules according to their similar characteristics (J. T. Yao, Vasilakos, & Pedrycz, 2013). In terms of coarse and fine relationship (Y. Yao, 2005, 2009), pixel granules are the refinement of the structure granules where every pixel granule is contained in the structure granular level.

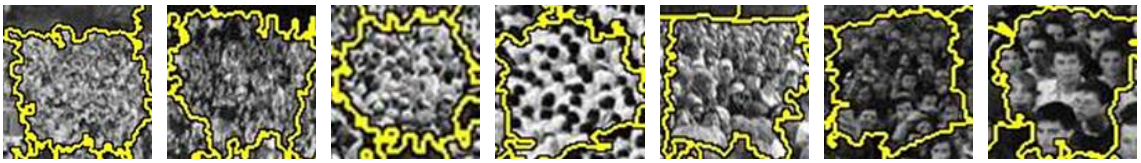


Figure 3.8: Sample dense crowd structure granules with variabilities in terms of illumination, scale of persons per area, perspective and inter-occlusion. Note that the scale of person per image area increases when view from left to right. Best viewed in color.

Structure granules are constructed by aggregating pixels (i.e. pixel granules) with similar structure feature vector, adapting the pixels clustering approach (Achanta et al., 2010) with refinement. The refinement is necessary in this work to enable auto-adaptability of structure granules to conform to the structure of local atomic regions. This is different from the existing cluster analysis solutions (Pedrycz & Bargiela, 2012; Bargiela, Pedrycz, & Hirota, 2004; Pedrycz & Bargiela, 2002; Tang & Zhu, 2013) that use distance measures such as the similarity between two granules defined as an average distance between sub-granules. More precisely, the proposed approach commences by initializing the number of structure granules, K, in an image, I. The greater the value of K, the finer is the crowd image partitioned, generating more structure granules. The initial structure granule centers, $\{c_k\}_{k=1}^{K}$, for an image, I, with N pixels is regularly seeded at a grid interval $G = \sqrt{\frac{N}{K}}$. Each c_k is represented by a feature vector, $\boldsymbol{\upsilon}_{c_k s} = \left(v_{c_k 1}, \ldots, v_{c_k s}, \ldots, v_{c_k S}\right)^{\top}$. Within the search region $(2G \times 2G)$ for each structure granule center, c_k, similarity of each feature, $v_{c_k s} \in \boldsymbol{\upsilon}_{c_k s}$ of structure granule center, c_k, with pixel, p, within the respective search region is defined as:

$$d_{ps}^{\tau} = \left\| v_{c_k s} - v_{ps} \right\|_2 \tag{3.1}$$

Anchor pixels for a structure granule are the pixels (i.e. pixel granules) that are associated with a specific structure granule center. The anchor pixels for each structure granule center, c_k are obtained by iteratively associating pixels in the image, I, to the nearest structure

granule center using the shortest pairwise distance. The pairwise distance measure, D^{τ}, is formulated as:

$$D^{\tau} = \sum_{s=1}^{S} \frac{d_{ps}^{\tau}}{m_s^{\tau-1}} \, , \, \tau \in \{1,2,3,\dots\} \tag{3.2}$$

$$\text{where } m_s^{\tau-1} = max(m_s^{\tau-2}, d_{mps}^{\tau-1}) \tag{3.3}$$

$$d_{mps}^{\tau-1} = max\left\{d_{ps}^{\tau-1}, \forall \; p \in 2G \times 2G\right\} \tag{3.4}$$

such that $d_{mps}^{\tau-1}$ is the maximum distance of a structure granule centre, c_k, with the pixels within the respective search region at iteration $\tau - 1$. The anchor pixels together with its respective structure granule center will form a structure granule (i.e. a region within yellow outlines as shown in Figure 3.1 (Middle)).

Note that, $m_s^{\tau-1}$ is a novel adaptive varying scaling parameter in the GrCS. This is in contrast to the constant scaling parameter scheme employed in (Achanta et al., 2010). Due to complex texture variations in crowd scenes, compactness of structure granules in terms of crowd and background boundary adherences is essential to provide an informative granulated view to comprehend scene context. Inspired by Zelnik-Manor and Perona (2004), in this work, at each iteration, τ, the selection of the scaling parameter, $m_s^{\tau-1}$ for each d_{ps}^{τ} is computed by studying the local structure of the anchor pixels with structure granule center, c_k from previous iterations (Eq. 3.3). Using a scaling parameter that honors the local structures of structure granule enables self-tuning of the pixel-to-granule center distances according to the local statistic of different features of the granule. The adaptive varying scaling parameters automatically find, at each iteration, the scales that enable high structure affinity of pixels within each structure granule and low structure

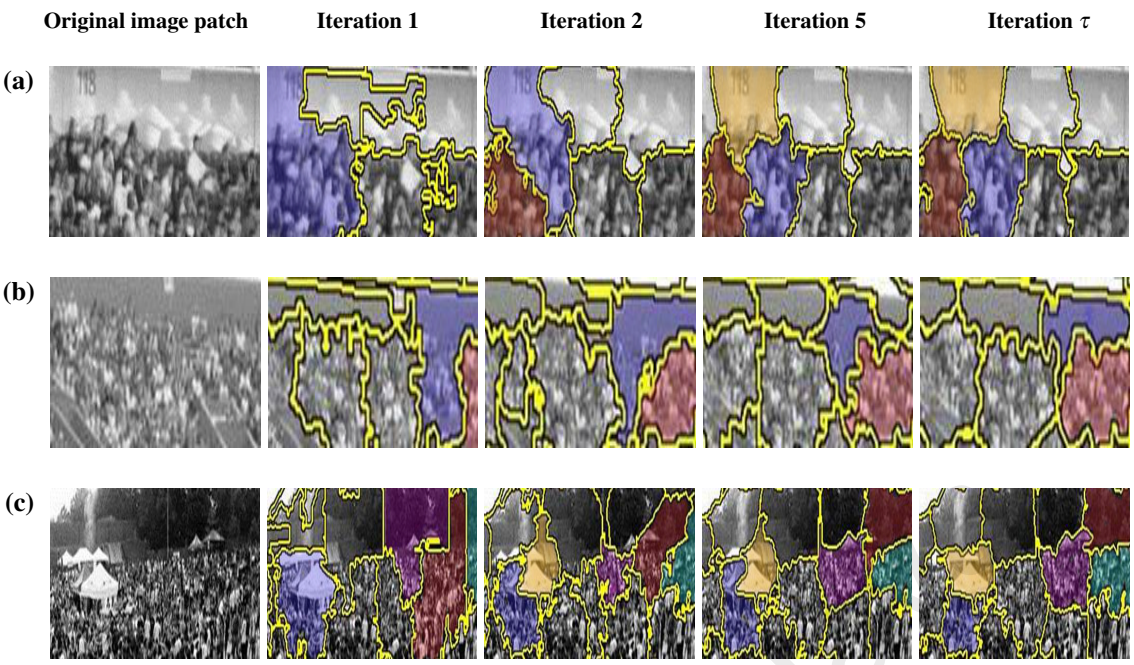


Figure 3.9: Transition of structure granules at each iteration. Structure granules with significant localization improvement are overlaid with different colors (i.e. purple, red, yellow, pink and green) to enhance the visualization of the improved separation between crowd and background regions over the iterations. (a) At iteration 1, it can be observed that the structure granule with purple overlay consists of crowd and background regions. After several iterations, at iteration τ, high localization of structure granules is achieved where crowd and background regions are well separated. That is, the structure granule with yellow overlay consists of background region only, whereas the structure granules with purple and red overlay consist of crowd region only. Similarly, (b) and (c) show the localization improvement of structure granules on two different crowd scenes. Best viewed in color.

affinity across neighboring granules for each structure feature, v_{ps} (as shown in Eq. 3.2). Section 3.3.4 demonstrates that this in turn facilitates distinct separation adhere to the natural boundaries between crowd and background regions in dense crowd images.

A set of new structure granule centers, $\{c_k^\tau\}_{k=1}^K \in I$ is defined at each iteration, where each c_k^τ is represented by the average of feature vector, $\boldsymbol{\upsilon}_{ps}$ of anchor pixels within the respective clusters. The optimized clusters constructed at this stage form a vocabulary of structure granules providing the granular description of the dense crowd image. Figure 3.9 shows examples of the transition of structure granules at each iteration. As the number of iterations, τ, increases, the localization of structure granules improves with optimized separation between crowd and background regions and eventually converges (see Appendix A for details). The pseudo code in Algorithm 1 describes the iterative

Algorithm 1 : Construction of structure granules

Require: An initial set of structure granule centers, $\{c_k\}_{k=1}^{K} \in I$, regularly seeded at a grid interval G and number of iterations, τ, where $\tau \in \mathbb{N}$

Ensure: A set of new structure granule centers, $\{c_k^{\tau}\}_{k=1}^{K} \in I$

```
repeat
  for each structure granule center, c_k do
    for each feature, v_ps do
      Compute d_ps^τ as to Eq. 3.1;
    end for
    Compute D^τ as to Eq. A.1;
  end for
  Associate pixels to the nearest structure granule center, c_k, by D^τ;
  Update set of structure granule centers, {c_k^τ}_{k=1}^K ∈ I;
until Separation between crowd and background regions is optimized
```

process to construct structure granules given the crowd scene image.

3.2.3 Crowd Segmentation

Given the structure granular, dense crowd segmentation task is posed as a classification problem. The aim is to achieve robust crowd regions inference by taking into consideration of the variability (as shown in Figure 3.7 and Figure 3.8) to infer class label (i.e. crowd or background) of input structure granules.

Random Forest (RF) is a term to describe an ensemble of decision trees. Unlike a single decision tree which is prone to bias to dominating class (Dietterich & Kong, 1995), RF is implemented due to the high generalization power yet able to avoid model overfitting, and being fast during training and testing (Breiman, 2001; Hoo, Kim, Pei, & Chan, 2014). Each random decision tree is generated by a random subset, $\mathbf{E}'$ of the labeled training structure granules with replacement. At a specific leaf node, the labeled training structure granules, $\mathbf{E}'_{node} = \{\mathbf{c}_i, l_i\}_{i=1}^{A}$ are recursively split into left, $\mathbf{E}'_{left}$ and right, $\mathbf{E}'_{right}$ node subsets, where $\mathbf{c}_i$ is a feature vector of structure granule, l_i is the corresponding class label (i.e. crowd or background) and A is the number of training samples. The splitting is

done given a set of thresholds, $\mathbf{T}$ and splitting function, f as:

$$\mathbf{E}'_{left} = \left\{ \mathbf{c}_i \in \mathbf{E}'_{node} | f(\mathbf{c}_i) < t \right\} \tag{3.5}$$

$$\mathbf{E}'_{right} = \mathbf{E}'_{node} \setminus \mathbf{E}'_{left} \tag{3.6}$$

At each leaf node, the threshold, $t \in \mathbf{T}$ that best split the training granules with maximized gain, ΔG is selected,

$$\Delta G = -\frac{\left|\mathbf{E}'_{left}\right|}{\left|\mathbf{E}'_{left}\right| + \left|\mathbf{E}'_{right}\right|} \cdot J_{left} - \frac{\left|\mathbf{E}'_{right}\right|}{\left|\mathbf{E}'_{left}\right| + \left|\mathbf{E}'_{right}\right|} \cdot J_{right} \tag{3.7}$$

where $J = -\sum_l p(l_i) \cdot (1 - p(l_i))$ is the Gini index and $p(l_i)$ is the class probability for l_i. Class labels of Q unseen structure granules, $\left\{\mathbf{c}_j\right\}_{j=1}^{Q}$ are inferred by traversing down all β decision trees. Each leaf node of a decision tree returns a prediction of the class label, l_j with class probability distribution $p(l_j|\mathbf{c}_j)$. The final class label (i.e. crowd or background) of structure granule is equated by averaging the probability estimate from each decision tree, defined as:

$$l_j^* = \arg\max_{l_j} \frac{1}{\beta} \sum^{\beta} p_\beta(l_j|\mathbf{c}_j) \tag{3.8}$$

The class labels of structure granules in an unseen image computed are used to infer the foreground (i.e. crowd) and background granules in the dense crowd scene image. The construction of foreground and background granules is a process of granulation. Such granulation process provides a granulated view of the image which is intended to be on par with the way a human would annotate crowd and background regions in a dense crowd scene.

3.3 Experiments

3.3.1 Dataset

Evaluations on the GrCS framework are conducted using 201 public benchmark datasets of real and synthetic dense crowd scenes obtained from (Idrees et al., 2013; Arandjelovic, 2008; Rodriguez, Sivic, et al., 2011; Fagette et al., 2014). These datasets consist of dense crowd scenes in various events, such as parades, concerts and rallies. The crowd in these datasets varies in terms of illuminations, crowdedness and perspectives. The resolutions of the images range from 240×320 to 1024×1024. To evaluate the efficiency of the proposed framework (i.e. conform precisely to the boundaries between crowd and background regions), the ground truth of crowd and background regions for real crowd scenes are manually annotated. Ground truth of each image is annotated at the pixel level, with careful labeling around complex boundaries of crowd. Examples of ground truth annotation are illustrated in the second row of Figure 3.13. The ground truth for synthetic crowd images is generated by the Agoraset crowd generator (Allain, Courty, & Corpetti, 2012). Each ground truth segment is highly accurate, i.e. adhering to the precise outline between crowd and background, where it would be almost infeasible to achieve manually (Courty, Allain, Creusot, & Corpetti, 2014).

3.3.2 Experiment Settings

In all the experiments, the number of structure granules, K is set to be 200 and the number of iterations, $\tau = 10$ which enables high localization of structure granules with adequate separation between crowd and background regions. The varying scaling parameter, $m_s^{\tau-1}$, for each d_{ps}^{τ} is initialized as $m_s^0 = 10$. Evaluation with different values of initialization constant generates consistent structure granules adhering to the boundaries of crowd. To construct granulated view of foreground (i.e. crowd) and background granules, random forest classifier is used with the number of random decision trees, $\beta = 2000$ and 100

randomly sampled variable at each split node. Dense crowd scene dataset is randomly divided into sets of 40 images to perform 5-fold cross-validation to avoid bias. Each structure granule is represented by the mean of feature descriptor, $\boldsymbol{\upsilon}_{ps}$ from pixel granulation, with entropy measures and pixel-wise SIFT (C. Liu, Yuen, & Torralba, 2011) features of anchor pixels and structure granule center, c_k. The feature responses of crowd and background structure granules are combined as input to train the random forest classifier.

3.3.3 Dense Crowd Segmentation

The effectiveness and robustness of the GrCS for real and synthetic crowd scenes understanding are demonstrated in the application of dense crowd segmentation. Evaluations are conducted by benchmarking the proposed framework with the multi-scale pixel grid approaches by Arandjelovic (2008) and Fagette et al. (2014). Each evaluation is compared against the benchmark dataset used in each respective approach.

Segmented crowd regions are shown as green overlay, whereas background regions with red overlay. For quantitative evaluation, the F-score measure is used according to the well-known PASCAL challenge (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010) to evaluate the accuracy of crowd segmentation by overlapping it with ground truth annotation (as per pixel basis).

Synthetic Crowd Scenes: Evaluations on synthetic dense crowd scenes are conducted to gauge the applicability of GrCS. Dense crowd segmentation on synthetic scenes is less taxing given the flat background texture. It is shown that when scales of person in crowd are uniform (as shown in row 1 of Figure 3.10), GrCS achieves similar or better F-score than Fagette et al. (2014) in classifying crowd and background regions. However, on crowd scenes with perspective distortion and varying crowdedness, GrCS is more

superior at discerning crowd and background regions, as illustrated in row $2-4$ in Figure 3.10. This is not the case for Fagette et al. (2014), where their segmentation does not accurately highlight the person in crowd. GrCS framework achieves good segmentation of individuals in crowd, simply because novel adaptive varying scaling parameter enables conformation of each structure granules adhering to the complex boundaries between crowd and background. With optimized structure granules, individuals in sparse crowd are adequately segmented.

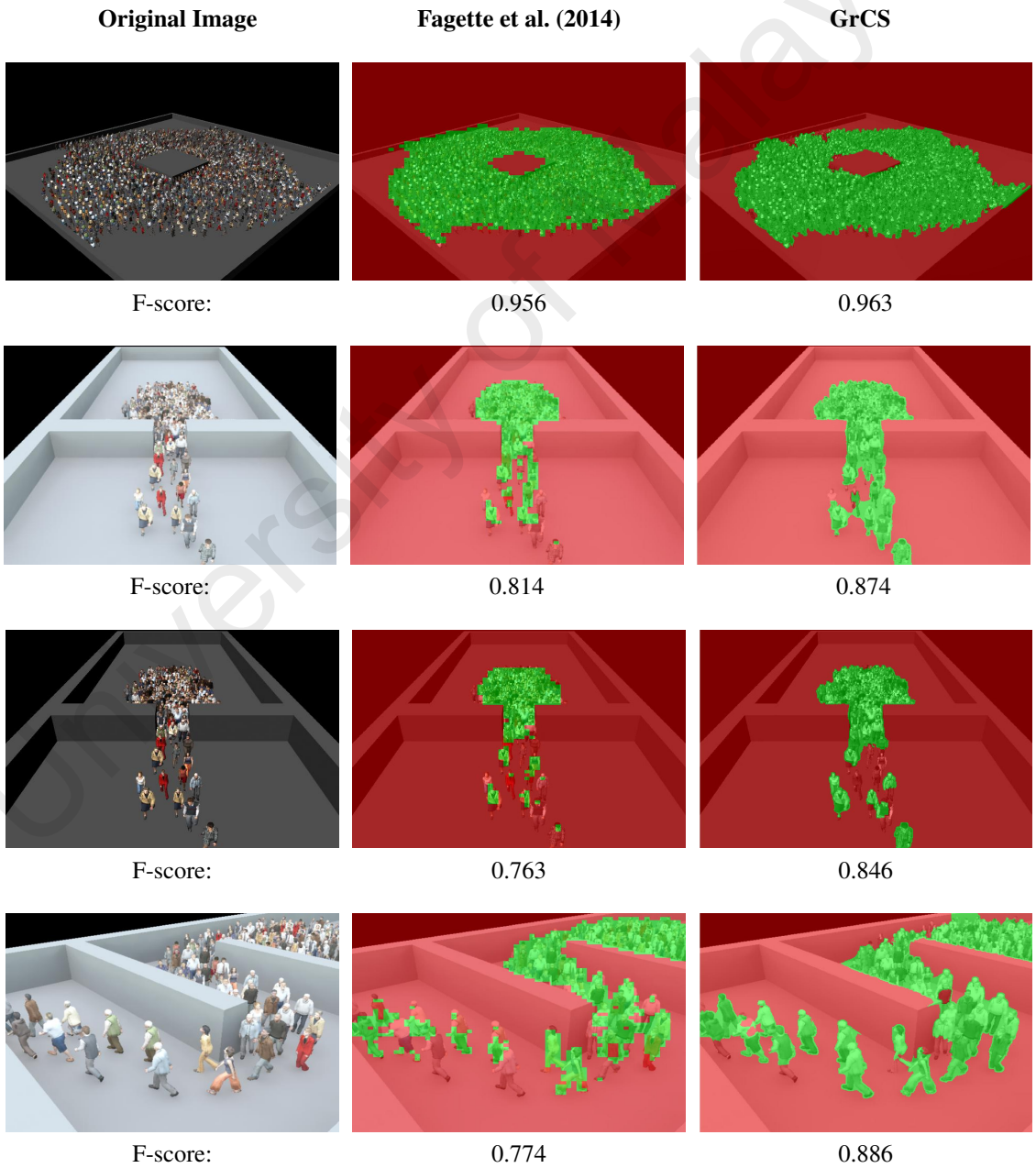


Figure 3.10: Comparative results of dense crowd segmentation on synthetic crowd scenes with Fagette et al. (2014). Best viewed in color.

Real Crowd Scenes: Contrary to synthetic scenes, real crowd scenes are more challenging given the varying crowd context, cluttered background and unconstrained physical layout of environment. The GrCS is further tested on real crowd scenes such as shown in Figure 3.11 and Figure3.12. Analogous with synthetic crowd scene, evaluation on real crowd scenes shows that when the scale of a person in a crowd are uniform where each person occupies only few pixels, the GrCS is comparable with Fagette et al. (2014) (as shown in row 1 of Figure 3.11). Evaluation on dense crowd scenes with perspective distortion and different crowdedness shows that the proposed method is able to cope better with varying scales of individuals in crowd to discern crowd and background regions in comparison to Fagette et al. (2014) and Arandjelovic (2008), as illustrated in row 3 of Figure 3.11 and row 2 of Figure 3.12. This is because the correlation among granules is exploited to represent structurally similar regions in crowd scenes and the variability of structures is taken into consideration during the granular information classification.

Background textures have significant influence on the crowd segmentation performance. For example in row 4 of Figure 3.11, Fagette et al. (2014) fails to segment crowd that has been overlaid by the steel barricades. Worst still, due to the crowd-like structure of steel barricade, it is mistakenly inferred as crowd segment. On the contrary, the GrCS is able to infer the actual crowd and background (i.e. steel barricade) segments. Arbitrary distribution of crowd and background regions is effectively outlined using GrCS (as shown in the fourth row of Figure 3.11 and the first row of Figure 3.12). It provides a more natural representation of crowd and background regions in comparison with Fagette et al. (2014) and Arandjelovic (2008). This essentially illustrates the advantage of granulation process that is adaptive to different crowd structure in scenes over pixel-grid. In addition, GrCS framework which utilizes Local Range of Intensity (LRI) feature is less susceptible to false segmentation.

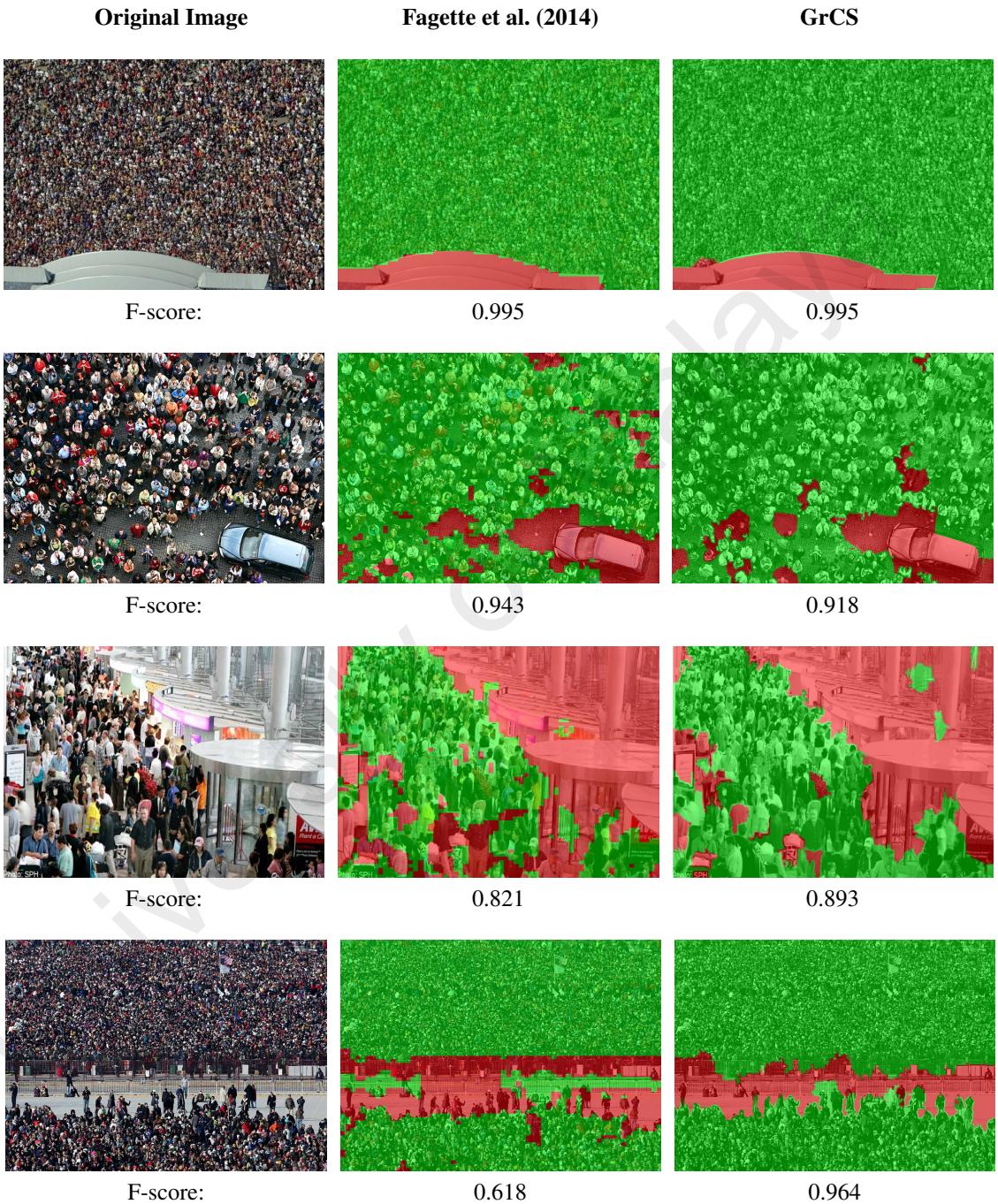


Figure 3.11: Comparative results of dense crowd segmentation on real dense crowd scenes with Fagette et al. (2014). Best viewed in color.

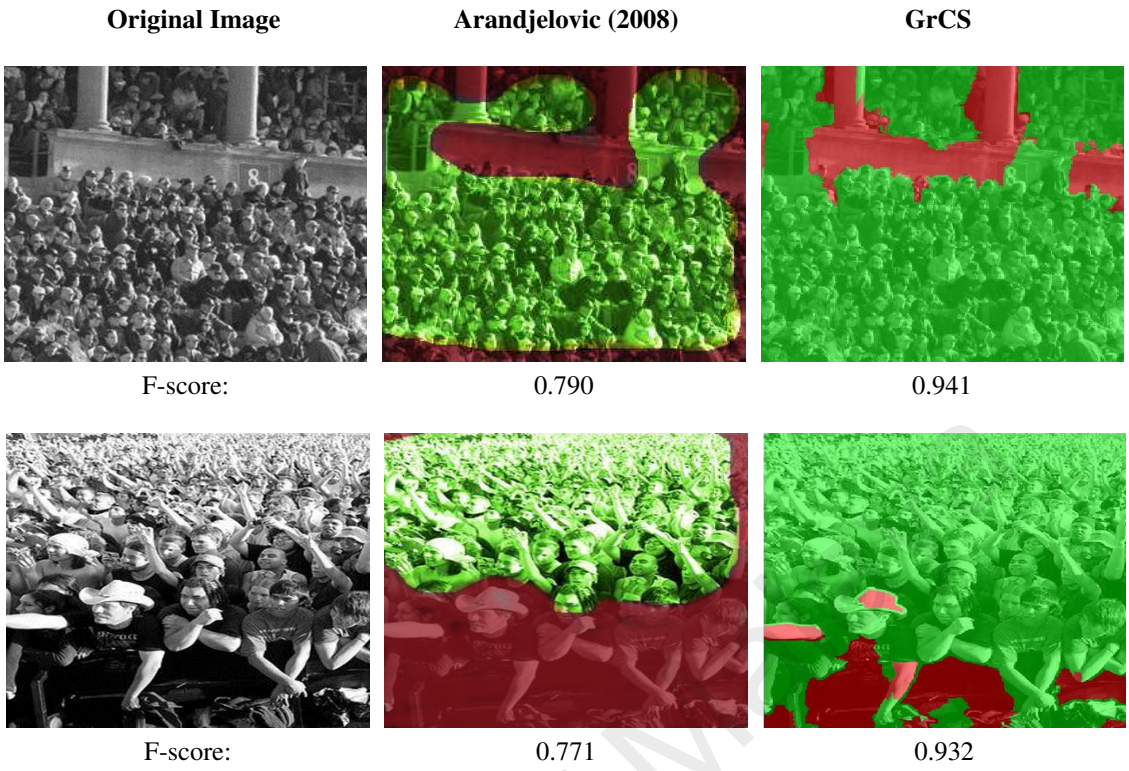


Figure 3.12: Comparative results of dense crowd segmentation on real dense crowd scenes with Arandjelovic (2008). Best viewed in color.

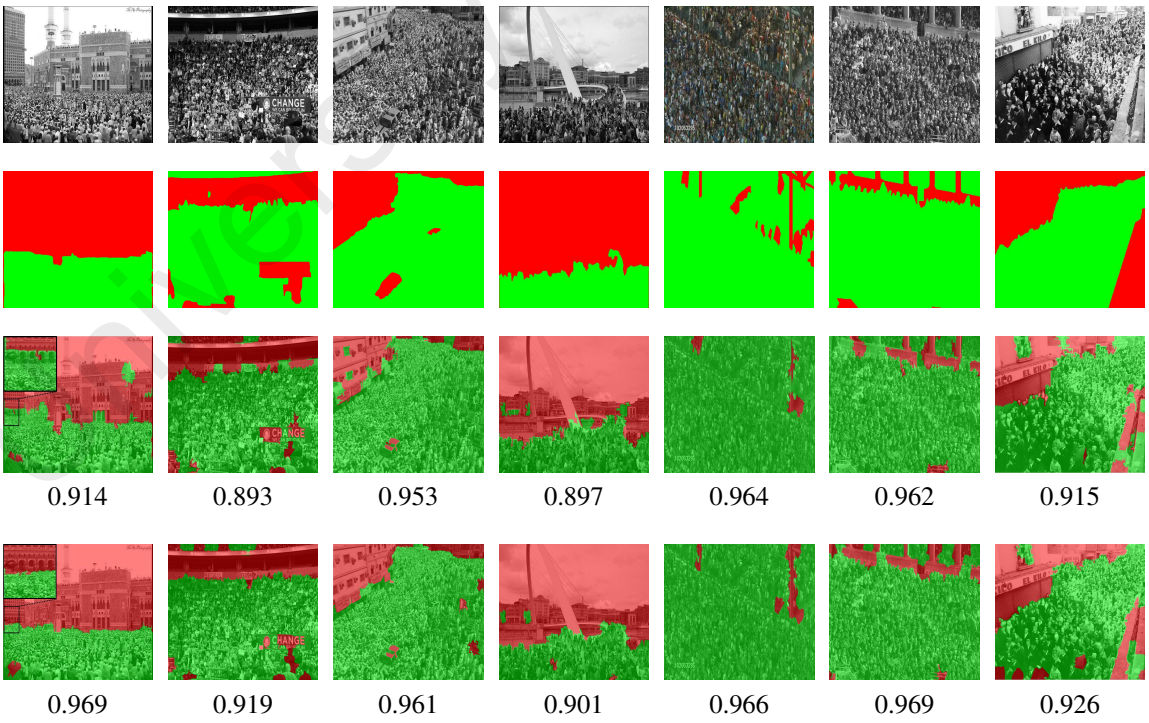


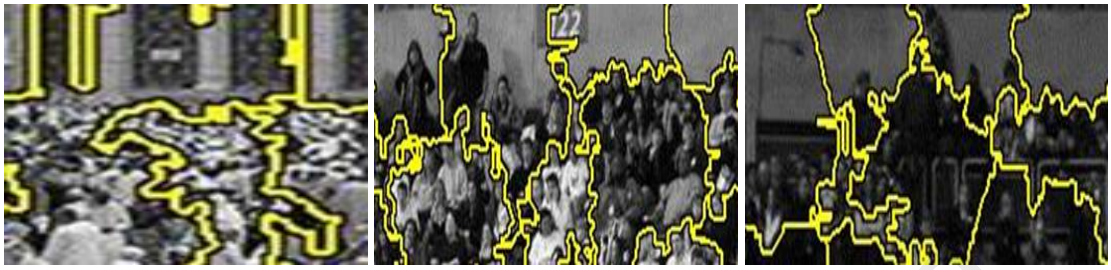
Figure 3.13: Comparative results of dense crowd segmentation on real dense crowd scenes with SLIC (Achanta et al., 2010). First row: real crowd scenes. Second row: ground truth annotations. Third row: crowd segmentation using SLIC (Achanta et al., 2010) with the respective F-score measures. Forth row: GrCS (adaptive varying scaling parameter) with the respective F-score measures. Best viewed in color.

3.3.4 Adaptive Varying Scaling Parameters

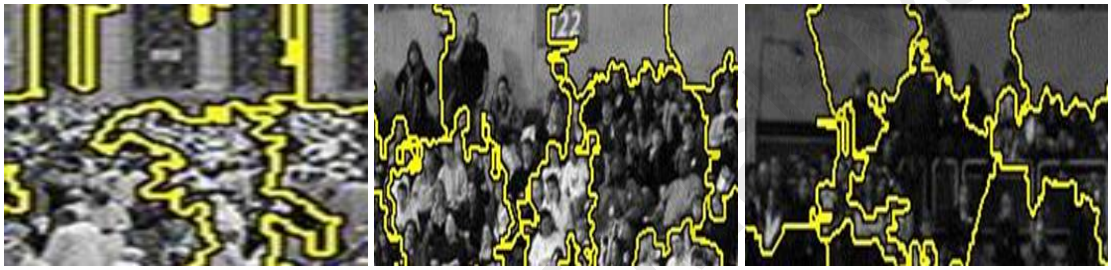
The proposed GrCS approach using adaptive varying scaling parameter is compared against constant scaling parameter by Achanta et al. (2010) on real crowd scenes. Examples of the ground truth and the segmentations results in comparison are shown in Figure 3.13. By using constant scaling parameter, crowd can be well separated from uncluttered background regions, but it performs poorly on complex and cluttered background. This is observed in the third row first column of Figure 3.13, where the ambiguous perimeter between crowd and building structure is inaccurately outlined. Moreover, as some of the structure granules constructed using approach by Achanta et al. (2010) contain both crowd and background texture (as shown in Figure 3.14), it is understandable that the granular information is prone to classification error. As illustrated in the first and second column of Figure 3.13, constant scaling parameter approach leads to textured regions of buildings inaccurately inferred as crowd, whereas the GrCS approach is able to define crowd and background regions corresponding to ground truth annotation.

To comprehend the influence of adaptive scaling parameter on dense crowd segmentation, Figure 3.14 provides visualization of the ground truth and the comparative results of structure granules using the novel adaptive varying scaling parameters and the constant scaling parameter by Achanta et al. (2010) (taken from random regions in crowd scenes from the first two columns in Figure 3.13). The results show that by using the constant scaling parameter (Achanta et al., 2010), the structure granules fail to adhere to the perimeters between different structures (particularly, crowd and background), in contrast to GrCS which uses adaptive varying scaling parameters. The main reason is, since each pixel, p, is represented by multiple structure features, v_{ps} that capture varying aspects of textures, so by using a constant scaling parameter for all d_{ps} throughout the iterations will not work well to capture the local affinity of each texture feature, v_{ps}, of pixels within the

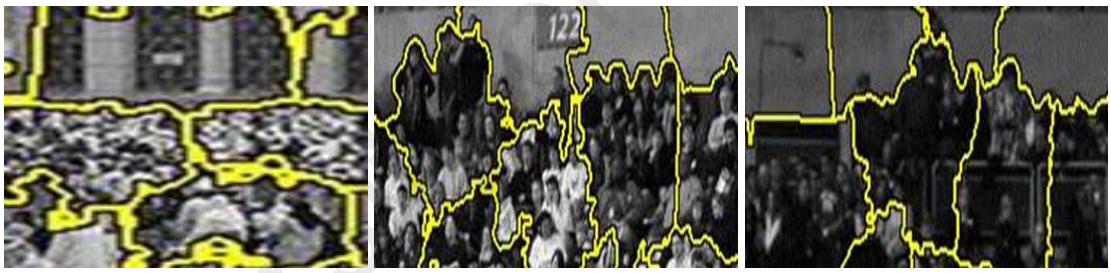
structure granule. Note that constant scaling parameter will act as normalization constant. Thus, any value of constant scaling parameter would generate similar structure granules, as shown in Figure 3.14a and Figure 3.14b.



(a) Constant scaling parameter = 10 (Achanta et al., 2010)



(b) Constant scaling parameter = 20 (Achanta et al., 2010)



(c) Proposed method, $m_s^{\tau-1}$

(d) Ground truth

Figure 3.14: Comparative results of structure granulation using constant value scaling parameter (Achanta et al., 2010) and the proposed adaptive varying scaling parameters. In ideal segmentation results, crowd regions are shown as green overlay, background with red overlay and blue line indicate ideal boundary between crowd and background. Boundaries between crowd and background of structure granules using adaptive varying scaling parameters are closer to the ground truth. Best viewed in color.

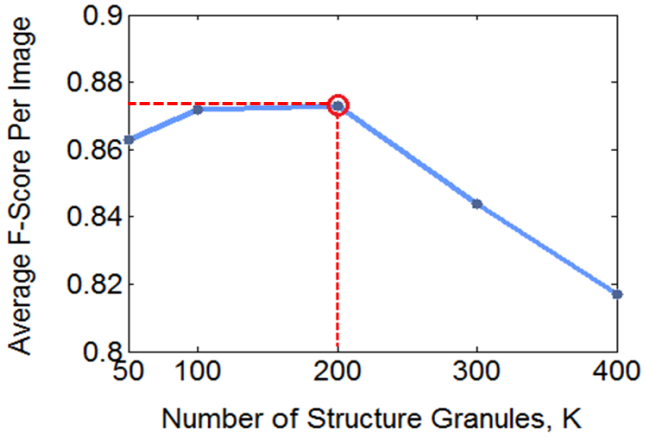


Figure 3.15: This figure shows analysis of average f-score measure per dense crowd image in terms of number of structure granules, K. For $K = 200$, the average f-score per image is 0.873.

3.3.5 Number of Structure Granules

The parameter K determines the number of structure granules in an image. The greater the K value, the more the structure granules constructed per image. Figure 3.15 provides the visualization of the influence of the parameter K on the crowd segmentation performance. The result shows that the higher the K value, the less precise is the segmentation per image. This is as expected, because with respect to the image size, with a greater K value, the image is decomposed into smaller size structure granules, where each granule contains fewer number of pixels. Consequently, fewer structures are present to infer the content (i.e. crowd or background) of the corresponding granule. Likewise, the smaller the K value, the fewer the structure granules constructed per image, which in turn generate larger size structure granules. When the size of a structure granule becomes too large, it can no longer represent the structure characteristics of a local region. In all the experiments in this chapter, K is empirically set to be 200 (Figure 3.15), which forms compact structure granules that outlines the natural boundaries between crowd and background regions.

3.3.6 Compactness of Structure Granules

Given the feature descriptor, $\boldsymbol{\upsilon}_{ps}$, of each pixel in a dense crowd scene, structure granules are formed by aggregating correlated pixel granules (detailed in Section 3.2.2). The sought after characteristics of structure granules are:

- Boundaries between the structure granules of crowd and background regions are distinct, with each segregated into different structure granules.
- Structure granules conform to the natural outline of arbitrary distribution of crowd.
- Each structure granule contains structurally similar pixels of dense crowd scenes (i.e. high localization accuracy). This is to cope with varying scales of individuals due to perspective distortion.

The intuition is that each structure granule provides a compact and localized primitive characterizing the local structure for dense crowd segmentation.

An example of the structure granules (pixels granulation) on dense crowd scene constructed using the GrCS is shown in Figure 3.16 with yellow outlines indicates the partitions between granules. It is observed that this dense crowd scene has severe perspective distortion of crowd. Still, the GrCS is able to aggregate neighboring individuals of similar scale into structurally uniform atomic regions. Groups of individuals in crowds that appear much bigger in the images are segregated into different granules from those that appear smaller (regions in orange, green and red box). At the same time, crowd regions with different crowdedness are observed to be grouped into separated granules. Despite complex background clutters (i.e. trees, building patterns and image noise), the aggregation of correlated pixels enables precise segregation of crowd and background regions, as illustrated in blue box.

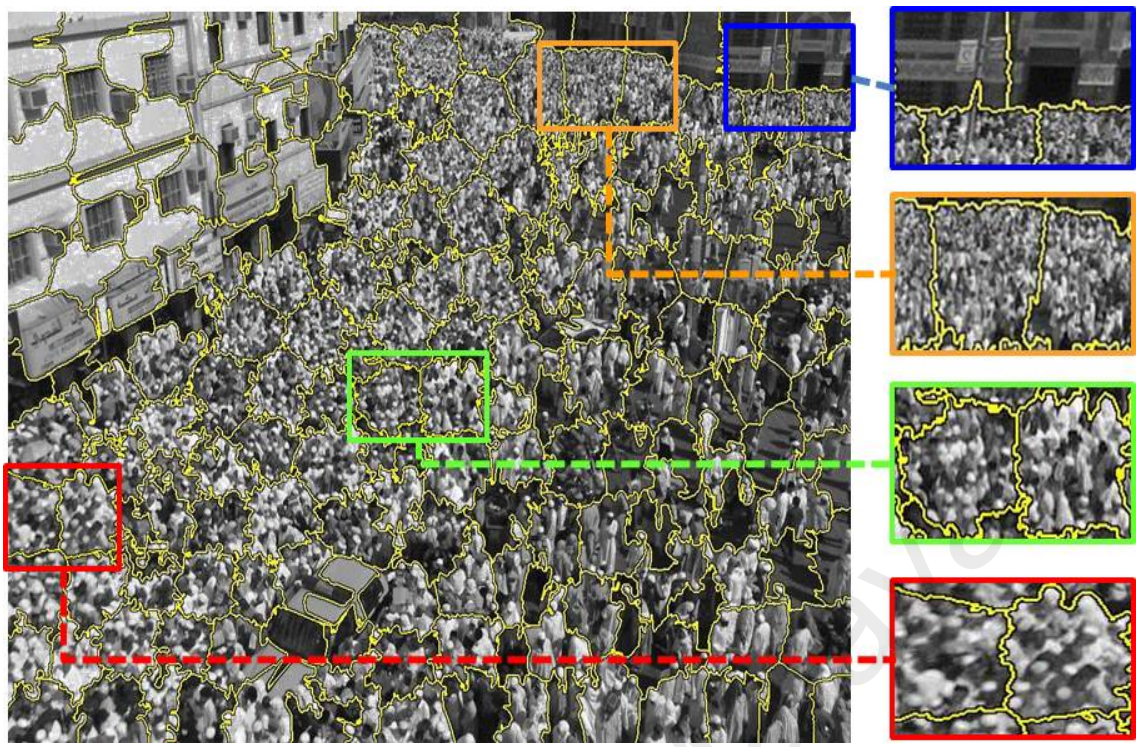


Figure 3.16: Examples of structure granules on a dense crowd images. Yellow outline indicates the partitions between granules. (Blue box) clear separation of structure granules between crowd and background. (Orange, green and red boxes) structure granules of crowd with significantly different crowdedness. Best viewed in color.

To evaluate the boundary adherences (compactness) of structure granules in the crowd scenes quantitatively, the local grouping of structurally similar pixels is considered as a clustering problem, where the widely adopted measurement in clustering evaluations (i.e. *Purity* (C. C. Aggarwal, 2004)) is used. The *Purity* measure of structure granules is utilized to quantify the quality of the granules against the pixel-level ground truth annotation labels (i.e. crowd or background). A structure granule is considered pure if it contains label from only one class, which is either crowd or background. Otherwise, a structure granule is considered as impure. In this context, an impure structure granule denotes that there is inaccurate separation between crowd and background regions. The accuracy of separation is quantified by the *Purity* measure, which is bounded within the $[0, 1]$ range. A higher *Purity* measure suggests a higher accuracy of boundaries between crowd and background regions.

Figure 3.17 shows the comparison and relative improvement of the structure granules

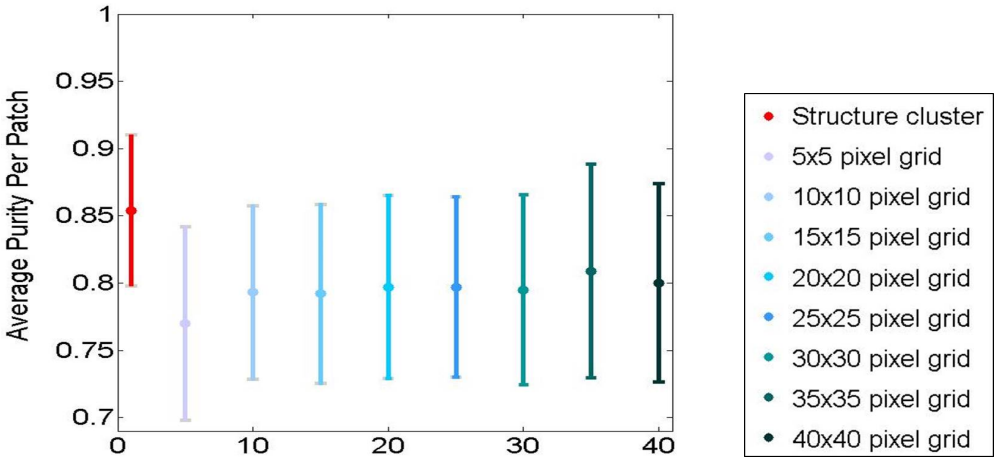


Figure 3.17: Quantitative comparison of the boundary adherence (*purity*) measure of structure granules with different pixel-grid sizes. Means are shown in dots, standard deviations with bars. Best viewed in color.

against varying scales of pixel-grid representation. Due to the aggregation of correlating pixel granules, structure granules are able to conform to the natural boundaries between different structures, in particular, crowd and background structure. Accordingly, the average purity measure of the proposed structure granules (0.854) outperforms the pixel-grid representation in all scales. Note that the proposed structure granules representation does not require manual intervention to achieve optimal boundaries adherence.

Furthermore, the *Purity* measures of structure granule per dense crowd scene is shown in Figure 3.18. It is observed that there are few dense crowd scenes with relatively lower *purity* measures. Upon scrutinizing the results, it is observed that these images correspond to poorly illuminated dense crowd scenes, i.e. concerts and cinema (as shown in Figure 3.19), in which the lack of illumination may weaken informative textures structures and diminish scene details. Even so, the *Purity* measures of the respective images are above 0.73.

3.4 Summary

This chapter has explored a new research direction in dense crowd scene analysis using the theory and principles of granular computing (GrC) to conceptualize dense crowd segmentation problem at different levels of granularity. Structure granules constructed by

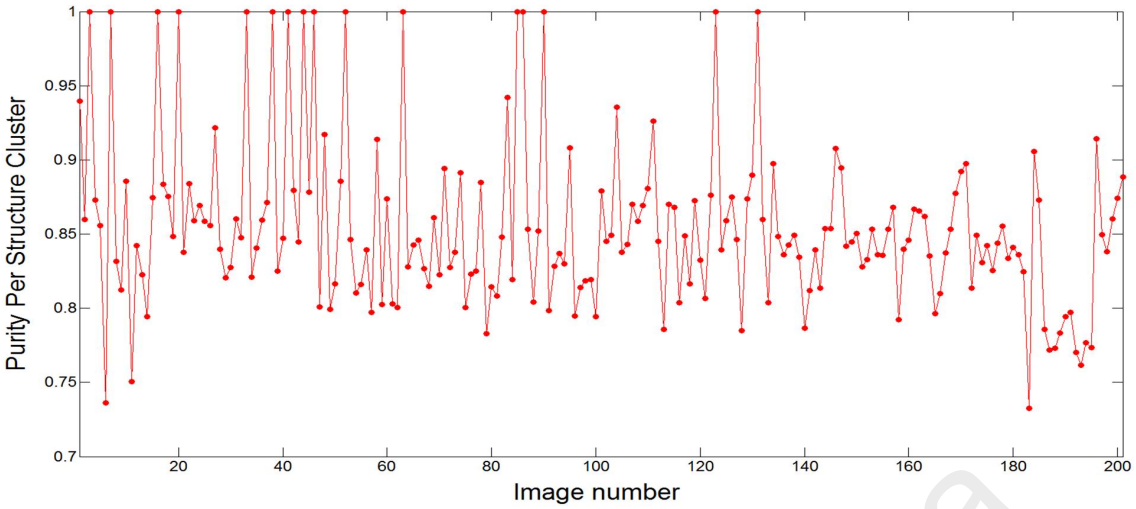


Figure 3.18: The boundary adherence (*purity*) measure per structure granule with respect to image. The average purity is 0.854.

Figure 3.19: Example dense crowd scene images with poor illumination. Lack of illumination may weaken informative textures structures and diminish scene details.

aggregating similar neighboring pixel granules are served as primitive characterizing local textures instead of regular pixel-grid. Experimental results on public and synthetic dense crowd scenes have shown that the granulation approach is effective in grouping structurally similar pixels into clusters to cope with perspective distortion, varying crowdedness and cluttered background for an effective interpretation of crowd and background regions.

Though the structure granular is effective in outlining boundaries between multi-scale crowd and background regions, the basis of granules for all granularity level are texture features. Thus, granulated view of different granularity level is limited when crowd scenes are poorly illuminated. Future investigation includes identifying texture features that are more robust towards characterizing poor illuminated crowd scenes.

CHAPTER 4: DENSE CROWD DENSITY ESTIMATION

Crowd density is one of the important aspects in dense crowd analysis when administering crowd well-being. Specifically, the number of individuals in a crowd can be an indicator of the comfort level in crowded scenes. It can also be a cue of imminent crowd disasters, i.e. crowd crush. Crowd disasters often occur when density in crowd become so great that individuals are crammed together. Physical forces from various directions cause individuals to fall, thus creating a domino effect that forces individuals to either step on each other or fall as well (Helbing et al., 2014). For instance, the recent crowd crush in the Shanghai 2014 New Year's Eve revelry which claimed 36 innocent lives.[1]

In some scenarios, having an accurate estimate of people count is important for historical record or to signify the legitimacy and effectiveness of a social movement, e.g., political rallies or protests. American sociologist Charles Tilly described that movement which demonstrates strength in the number of participant, as part of the WUNC (worthiness, unity, numbers and commitment) display, is a vital form of claim-making and measure of its success (Tilly, 1999). Thus, for many events, crowd size is a contentious issue. An accurate empirical estimate is required to prevent dispute of crowd count results from different parties (i.e. media, oppositions, organizers, etc.). The 1995 Million Man March is an example of a large crowd density estimate dispute between the organizer and the police (McPhail & McCarthy, 2004).

Despite the importance of keeping track of crowd density, as mentioned in Section 2.3, employing professional to estimate crowd size is infeasible. This is predominantly due to the sheer number of individuals in an unconstrained dense crowd environment. Thus, in this chapter, a novel algorithm is introduced for regression-based dense crowd density estimation. On the contrary to existing methods (Idrees et al., 2013; K. Chen et al.,

[1]BBC News: http://www.bbc.com/news/world-asia-china-30646918

2012), the proposed algorithm partitions images into irregular size granules conforming to the image context for density estimation. The preceding chapter has demonstrated the importance of studying the correlation among image granules at different levels of granularity in outlining natural boundaries between crowd and background (non-crowd) regions. The structurally meaningful atomic regions (i.e. granules) can serve as primitive regions to extract features for density estimation. The aim is to carry out reliable low-level feature extraction to infer accurate density of individuals in dense crowd scenes.

The remainder of this chapter is organized as follows: Section 4.1 describes the motivation of the proposed density estimation approach, followed by the formulations of the dense crowd density estimation strategy in Section 4.2. Section 4.3 presents and discusses the experimental results. The proposed density estimation approach is evaluated using public dense crowd dataset. Finally, conclusions are drawn in Section 4.4.

4.1 Dense Crowd Density Estimation

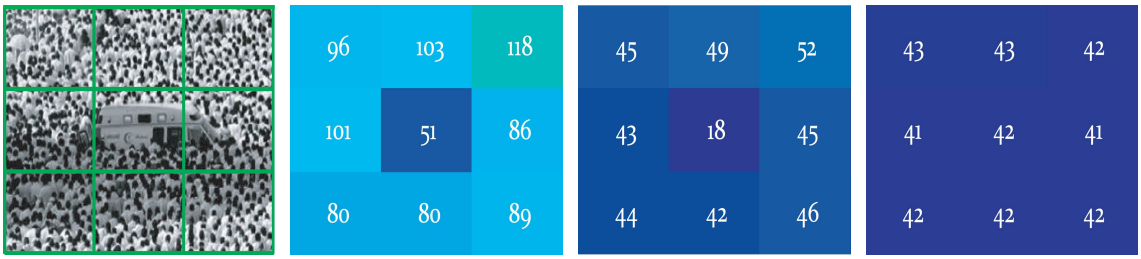
While object detection research in the field of computer vision has been improved significantly over the recent years, analyzing dense crowd scenes (particularly, density estimation) remains challenging (Rodriguez, Laptev, et al., 2011). This is because dense crowd scenes are characterized by the co-occurrence of a large number of individuals gathered closely together. The complexity often manifested itself in the frequent, partial or complete occlusion between individuals (Ali et al., 2013).

Since delineating individuals in dense crowd scenes are difficult (because of the spatial overlaps), most existing density estimation approaches (Marana et al., 1998; Chan et al., 2008; Chan & Vasconcelos, 2012; Idrees et al., 2013; K. Chen et al., 2012) obviates the steps to detect and / or track individuals. They put emphasis on extracting a set of low-level image feature. This paradigm of density estimation is based on regression, where the relationship between the extracted features and the density of individuals is

Figure 4.1: Example dense crowd scenes with perspective distortion. Individuals who are closer to the camera view appear larger than those who are positioned further away from the camera.

learned. However, a problem commonly encountered in regression based approach is perspective distortion, where individuals who are closer to the camera view appear larger than those who are positioned further away from the camera (as illustrated in Figure 4.1). The problem is exacerbated when single regression function is used for the whole image space. To address this problem, perspective normalization plays a key role by bringing the perceived size of individuals at different depths to the same scale (Loy et al., 2013). Another approach is to divide the image space into different pixel-grids and each pixel-grid is modeled by a regression function to mitigate the influence of perspective distortion. Such approaches rely on local features modeling through the analysis of pixel-grids (Ma, Huang, & Liu, 2010; K. Chen et al., 2012; Idrees et al., 2013).

Despite the promising results of density estimation using pixel-grid approaches (Idrees et al., 2013; K. Chen et al., 2012), it is susceptible to the constrain of pixel-grid. That is, conformations to the natural outline between crowd and background (non-crowd) are difficult to achieve (see Figure 4.2a). Consequently, one can observe in Figure 4.2d that imprecise delineation of crowd and non-crowd regions, as well as assumption of dependency between pixel-grids can lead to inaccurate estimation of person count. This is because extracted features are not characterizing either crowd or background only. It is worth noting that assuming dependency between granules is impractical since fundamentally crowd density and distribution varies from regions to regions in unconstrained



(a) Pixel-grids **(b)** Ground truth count **(c)** Estimated count (Before MRF) **(d)** Estimated count (After MRF)

Figure 4.2: Dense crowd density estimation by Idrees et al. (2013). The dependency between pixel-grids is modeled by multi-scale MRF to enhance density estimation. Green outline indicates the partitions between pixel-grids. Crowd density for pixel-grids consisting of crowd and background (i.e. non-crowd) regions have been estimated to have similar density with crowd-only pixel-grids after dependency modeling. Best viewed in color.

public scenes. For instance, background elements can be randomly positioned within dense crowds, as shown in Figure 4.2a.

It is thus necessary to partition dense crowd scenes into granules that conformed to the natural outline between crowd and non-crowd regions. The granular computing based approach (described in Chapter 3) is extended to allow estimation of crowd density without tracking or segregation of individuals. Importantly, in contrast to a pixel-grid based approach (Ma et al., 2010; K. Chen et al., 2012; Idrees et al., 2013), the atomic regions (i.e. structure granules) can serve as meaningful primitive regions to extract features essential for density estimation. This strategy is applicable to density estimation in public dense crowd scenes, i.e. scene-invariant.

In addition, the proposed approach is motivated by the fact that no single feature can provide sufficient information for density estimation in dense crowd scenes. As noted by Idrees et al. (2013), this is predominantly due to low resolutions imagery, perspective distortion and severe occlusions (detailed in Section 1.2). One can, however, observe that dense crowds portray textures which can be employed to infer crowd density. There is a relationship between low-level features and crowd density that is expected to facilitate dense crowd density estimation (Marana, Velastin, Costa, & Lotufo, 1997).

4.2 Proposed Dense Crowd Saliency Detection Framework

Given a dense crowd image, the aim of this work is to estimate the number of individual in the image. In a public scene, the density of individuals can varies from region to region. This density variation is mainly due to the effects of perspective distortion (see Figure 4.1) or constraints imposed by the environment layout. Thus, the proposed approach commence by representing structure granules using texture features. Note that the structure granules were formed from the aggregation of pixels with similar feature, described in Chapter 3. This is to facilitate in distinguishing between crowd and background (i.e. non-crowd) regions for density estimation. Crowd regions with different coarseness are also represented with different granules (as shown in Figure 3.16). Therefore, unlike existing density estimation approach, the proposed approach does not assume similarity of density in adjacent granules (i.e. dependency between granules).

Formally, a dense crowd image, $I = [\boldsymbol{\upsilon}_{gs}] \in \mathbb{R}^{G \times S}$, where G is the number of granules in an image and S is the number of features for each granule, g. Each granule, g, in a dense crowd image, I, is represented as a feature vector, $\boldsymbol{\upsilon}_{gs} = \left(v_{g1}, \ldots, v_{gs}, \ldots, v_{gS}\right)^{\top} \in \mathbb{R}^{G \times S}$, where $g = \{1, \ldots, G\}$ and $s = \{1, \ldots, S\}$. The feature vector, $\boldsymbol{\upsilon}_{gs}$ is formed by the mean of feature descriptor of each pixel, p, within the respective granule. The feature descriptor for each pixel, $\boldsymbol{\upsilon}_{ps}$, is the concatenation of S different and complementary features. The texture features used in the proposed approach to represent pixels are discussed in the following subsection. Dense crowd density estimation problem is subsequently formulated as a regression problem. In particular, a mapping function between feature vectors input and a scalar-valued crowd density output is learned.

4.2.1 Granular Representation of Dense Crowd Images

Although dense crowd can be irregular at a coarse level, the texture of crowd tend to correspond to a harmonic pattern (i.e. regular texture) at a finer scale patches (Idrees et

al., 2013), such as pixel-grids or granules. Moreover, crowd regions tend to present large number of texture features. As one can observed from Figure 4.1, this is because of the appearance variations of crowd. These texture features carry strong cues regarding the number of people in a scene (Loy et al., 2013). Thus, crowd regions in these patches can be treated as texture for processing.

In this work, dense crowd images are represented as structure granules for density estimation. It is the basic aspect of dense crowd scenes in this work, characterizing structurally meaningful atomic regions that distinguish between crowd and background regions for low-level feature extraction. The texture feature vector for each granule is the mean of texture features of pixels within the respective granule. The texture feature vector from each granule are used as description of the crowd, where a direct mapping between the features and crowd density is learned. In the proposed approach, the texture of dense crowd scenes is represented by the Local Standard Deviation (LSD), Dense Scale-Invariant Feature Transform (DSIFT) and Phase Congruency (PC). The proposed framework is, however, not restricted to these sets of features employed in this chapter. Diverse sets of features can be exploited to enhance and adapt to various dense crowd analysis researches.

4.2.1.1 Local Standard Deviation(LSD)

The proposed approach is inspired by the fact that dense crowd regions with different density tend to generate distinct local texture patterns, as shown in Figure 4.3. That is, highly dense crowd regions (as shown in the first column of Figure 4.3) comprise of fine patterns, whereas moderately dense crowd regions (as shown in the third column of Figure 4.3) mostly contain coarse pattern. As related by Davies et al. (1995) and Marana et al. (1998), there is a correlation between crowd density and edge feature of crowd. Accordingly, this proposed approach is motivated to use edge feature to characterize crowd

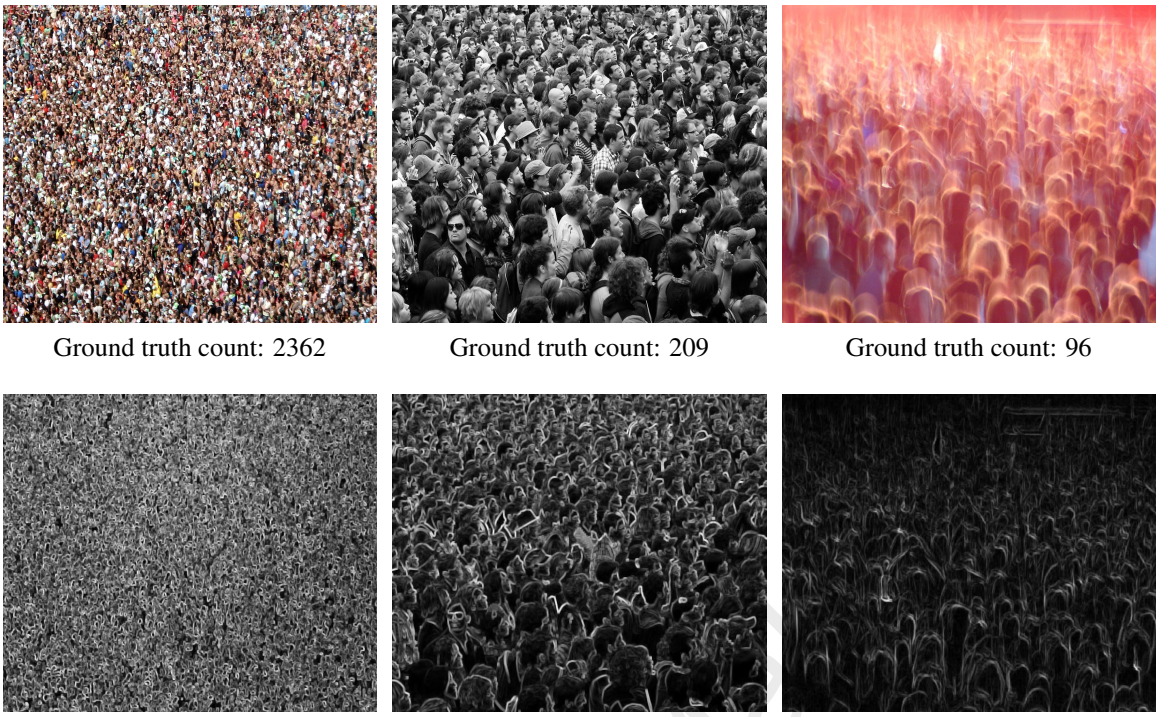


Figure 4.3: (Top row) Example dense crowd scene images. (Bottom row) Images of local standard deviation (LSD) using 5×5 neighbourhood. Note that the crowd density decreases when view from left to right. Best viewed in color.

regions.

To this end, Local standard deviation (LSD) is employed to capture the local image structure, i.e. edges, formed by mass of crowd in dense crowd images. This is because LSD is a computationally simple and practical edge detection mechanism (Lloyd, 2006). The output of LSD is a measure of the local average contrast. Specifically, calculating the LSD of pixels in a neighborhood can indicate the degree of variability of pixels intensities in that local region. Strong intensity contrast / variability of pixels characterize edges in images.

Given a dense crowd image, LSD calculate the standard deviation of pixel intensities in a 5×5 neighborhood centering each pixel of interest (i.e. all the pixels in the image). The output of LSD is assigned to the respective pixel of interest. One of the main advantages of using LSD in the proposed approach is that edge sharpness of crowd images can be quantified. This is essential to delineate the various texture features in dense crowd

images for density estimation.

4.2.1.2 Dense Scale-Invariant Feature Transform (DSIFT)

DSIFT (Vedaldi & Fulkerson, 2010) is a variation of the SIFT algorithm (Lowe, 2004), which is a state-of-the-art keypoint based approach to characterize local gradient information. By using SIFT, the number of interest points extracted from an image varies based on the image content, making the information incorporation on spatial configuration complicated (Tuytelaars, 2010). Conversely, DSIFT extracts SIFT histogram for all pixels with overlapping patches. Compared to sparse features (e.g. SIFT (Lowe, 2004), interest points (Mikolajczyk & Schmid, 2004)), dense features results in a good coverage of the entire scene (Tuytelaars, 2010). This produces a constant amount of features per image area that contain essential information of the image content.

As one can observed in crowd regions with highly irregular repetitive grain (as shown in Figure 4.3 (Top row)), it is likely to have similar texture element around different regions of crowd, formed by parts of peoples (Idrees et al., 2013). The local intensity gradient can reveal local individual appearance, such as head and shoulder, which is informative for density estimation (Loy et al., 2013). Therefore, in addition to edge feature of crowds, DSIFT is used in the proposed approach to model the appearance cue of crowd. DSIFT algorithm is implemented to extract feature descriptor for each pixel in a dense crowd image. The DSIFT feature descriptor corresponds to the spatial coordinate of image pixels, forming a dense description of the image.

Given a dense crowd image, the feature descriptor of each pixel of interest is constructed by overlying a window centering the pixel of interest. Each local window is further divided into smaller sub-windows (e.g. 4×4) where gradient orientation and magnitude is quantized into an 8 bin histogram in each sub-window. The feature descriptor of the pixel of interest is formed by concatenating the histogram of sub-windows,

obtaining a $4 \times 4 \times 8 = 128$ dimensional vector as the SIFT representation.

4.2.1.3 Phase Congruency (PC)

The gradient-based texture features, i.e. LSD and DSIFT, are sensitive to image illumination variations (Kovesi, 1999). Hence, these extracted features can be image dependent. To compensate and complement the set of features used to represents textures of structure granules, a dimensionless measure of feature significance that is invariant to image illumination is desired. Such measure can provide absolute quantifications of feature significance that is applicable to any dense crowd scene images.

Studies by Oppenheim and Lim (1981) have shown that phase information of images can retain the important features of image context. Interestingly, the Local Energy Model developed by Morrone and Owens (1987) postulates that features can be perceived at spatial positions of maximum phase congruency within an image in the frequency domain. Hence, the advantage of this model is that it is not based on local intensity gradient for feature detection. These texture features detected include edges and lines.

To construct a dimensionless measure of phase congruency of dense crowd images that is invariant to image illumination, the proposed approach uses the method introduced by Kovesi (1999). Kovesi (1999) scheme calculates the phase congruency with Log-Gabor wavelet filters (Field, 1987), which work as bandpass filters. It allows arbitrary large bandwidth filters to be constructed while maintaining a zero DC component in the even-symmetric filter (Kovesi, 2000). Hence, the phase congruency of a pixel p in a dense crowd image, I, is express as the summation over orientation o and scale n:

$$PC(p) = \frac{\sum_o \sum_n W_o(p) \lfloor A_{no}(p) \Delta\Phi_{no}(p) - T_o \rfloor}{\sum_o \sum_n A_{no}(p) + \varepsilon} \tag{4.1}$$

$$\text{where } \Delta\Phi_{no}(p) = cos(\phi_{no}(p) - \overline{\phi}_o(p)) - sin(\phi_{no}(p) - \overline{\phi}_o(p)) \tag{4.2}$$

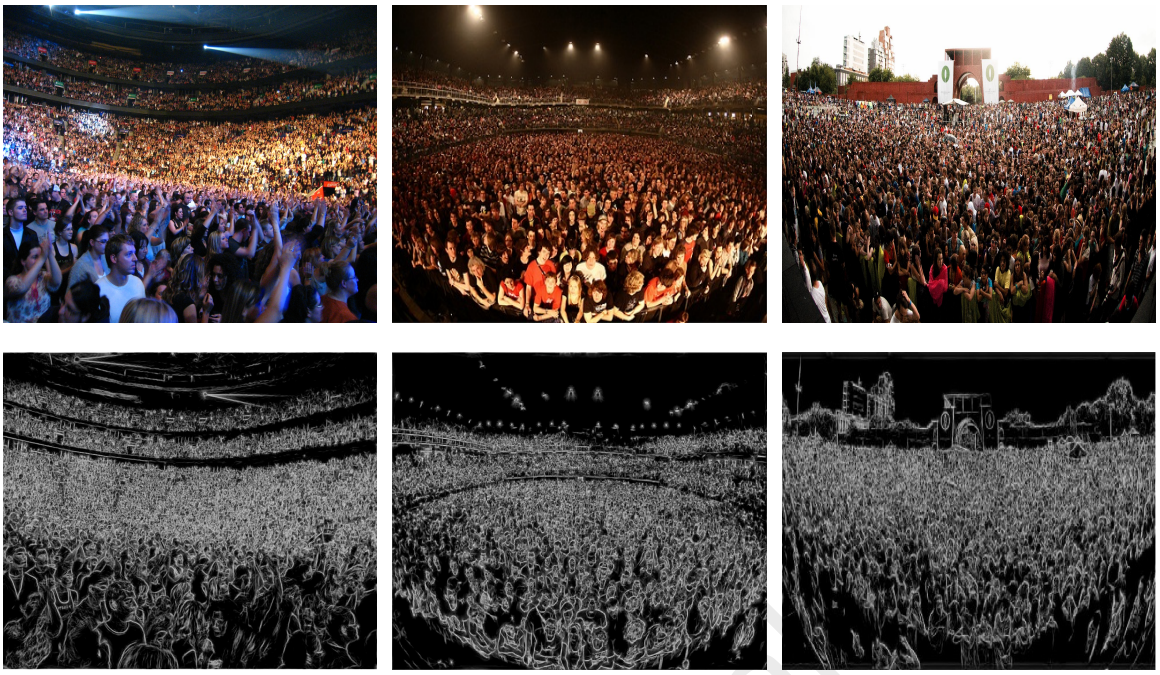


Figure 4.4: (Top row) Example dense crowd scene images. (Bottom row) Images of phase congruency (PC) corresponding to number of orientation $o = 6$ and scale $n = 3$. Note that the texture features is invariant to changes in illumination. Best viewed in color.

such that $\lfloor \cdot \rfloor$ is a floor function which denotes that the enclosed quantity is not permitted to be negative; $W_o(p)$ is a weighting factor based on frequency spread; $A_{no}(p)$ is the local amplitude of pixel p on scale n and orientation o; T_o is introduced to compensate for noise influence. A small denominator $\varepsilon = 0.0001$ is added to avoid division by zero (Kovesi, 2000). $\Delta\Phi_{no}(p)$ is a sensitive phase deviation measure, where $\overline{\phi}_o(p)$ is the mean phase angle for pixel p.

The output of the phase congruency takes on the values between $[0,1]$, providing an illumination invariant measure of texture features in dense crowd images. Figure 4.4 shows sample phase congruency outputs of dense crowd images.

4.2.2 Density Estimation by Regression

The texture feature vector, $\boldsymbol{\upsilon}_{gs}$, of a structure granule in a dense crowd image is the mean of feature descriptor of each pixel, p, within the respective granule. The feature descriptor, $\boldsymbol{\upsilon}_{ps}$, of each pixel, p, is the concatenation of the Local Standard Deviation

(LSD), Dense Scale-Invariant Feature Transform (DSIFT) and Phase Congruency (PC) texture features. Given the structure granules of dense crowd images, dense crowd density estimation task is posed as a regression problem. The aim is to learn the relationship between the texture features and the crowd density, for dense crowd density estimation of new scenes.

For sparse crowd scenes (as shown in Figure 1.2a) where lower crowd density and fewer occlusions among individuals are observed, linear regressor (e.g. ridge regression (Hoerl & Kennard, 1970)) may suffice. This is because the mapping between the features and people count typically presents a linear relationship (Loy et al., 2013). Nonetheless, given dense crowd environment, such as the scenes analyzed in this thesis, where there are severe partial and complete occlusions among individuals, a nonlinear regressor is required to capture the nonlinear trend in the feature space (Chan & Dong, 2011).

Formally, given M training data, represented as $\{\mathbf{x}_i, y_i\}_{i=1}^{M}$, $\mathbf{x}_i$ is the feature vector of structure granule, $\boldsymbol{\upsilon}_{gs}$, and y_i is the corresponding crowd density of the respective structure granule. The objective of regression is to predict the value of y given a new value of $\mathbf{x}$. In the proposed approach, the mapping between the texture features and the crowd density is estimated by learning a nonlinear function, in particular, a Kernel Ridge Regression (KRR). KRR with Radial Basis Function (RBF) kernel is employed owing to its promising performance in the literature for crowd density estimation (K. Chen et al., 2012; K. Chen, Gong, Xiang, & Loy, 2013). In its simplest form, a ridge regression function (i.e. $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$) is a linear regressor with a cost function as follows:

$$C(\mathbf{w}) = \frac{1}{2}\sum_i (y_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i))^2 + \frac{1}{2}\lambda \|\mathbf{w}\|^2 \tag{4.3}$$

where $\frac{1}{2}\lambda \|\mathbf{w}\|^2$ is a regularization term to avoid over-fitting of the training data. The parameter $\lambda > 0$ is determined via cross-validation. The model parameter $\mathbf{w}$ is determined

by minimizing the cost function $C(\mathbf{w})$.

The nonlinear version of the ridge regression, i.e. KRR, can be achieved via kernel trick (Shawe-Taylor & Cristianini, 2004). That is, constructing the ridge regression model in higher dimensional feature space induced by a kernel function. In this work, the RBF kernel function is used.

$$k(\mathbf{x},\mathbf{x}') = exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}) \tag{4.4}$$

where the kernel width parameter σ is determined via cross-validation. The KRR functions is given by:

$$f(\mathbf{x},\boldsymbol{\alpha}) = \sum_i \alpha_i k(\mathbf{x},\mathbf{x}_i) \tag{4.5}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_i\}$ are Lagrange multipliers used to solve the KRR minimization problem (Loy et al., 2013).

The estimated density of an unseen dense crowd image is the summation of the estimation obtained for all structure granules in the corresponding image.

4.3 Experiments

The following sections describe the dataset used in the experiments, experimental setup and dense crowd density estimation results.

4.3.1 Dataset

Evaluations on the proposed density estimation approach are conducted on public dataset obtained from (Idrees et al., 2013). This dataset consist of 50 dense crowd images collected mainly from Flickr[2]. The number of individuals in these images ranges between 96 and 4628, with an average of 1280 individuals per image. The scenes in these im-

[2]Flickr - Photo Sharing!: https://www.flickr.com/

Ground Truth count: 4628
(Most GT count)

Ground Truth count: 1940

Ground Truth count: 96
(Least GT Count)

Figure 4.5: Sample dense crowd images from the dataset with their corresponding ground truth count. The left and right images show the most and least ground truth count, respectively. ((Idrees et al., 2013))

ages are diverse, depicting dense crowd in various set of events, such as pilgrimages, concerts, marathons, stadiums and rallies. The ground truth provided are manually annotated, where the position of individuals are marked with a dot. Figure 4.5 illustrate some example of dense crowd images with the associated ground truth counts.

4.3.2 Experiment Settings

In the experiments, dense crowd dataset is randomly divided into sets of 10 to perform 5-fold cross-validation to avoid bias. Each image is represented by an approximately $G = 200$ granules.

4.3.3 Evaluation Metric

The performance of density estimation approach can be assessed by the similarity between the actual count and the estimated count of individuals in a scene (Idrees et al., 2013). Accordingly, the evaluation metric known as Absolute Difference (AD) is employed to quantify the density estimation results.

$$AD_i = \left| \zeta_i - \hat{\zeta}_i \right| \tag{4.6}$$

where i denote the ith patch (i.e. granule for the proposed approach) or image, ζ_i is the

Table 4.1: Comparative results of dense crowd density estimation with Idrees et al. (2013), Lempitsky and Zisserman (2010) and Rodriguez et al.(2011) using mean and standard deviation of Absolute Difference (AD) from ground truth. The proposed approach outperforms the state-of-the-art approaches.

Approach	AD Per Patch	AD Per Image
Rodriguez et al. (2011)	-	655.7 ± 697.8
Lempitsky et al. (2010)	-	493.4 ± 487.1
Idrees et al. (2013) - before MRF	10.2 ± 18.9	468.0 ± 590.3
Idrees et al. (2013) - after MRF	-	419.5 ± 541.6
Proposed	$\mathbf{6.4 \pm 6.6}$	$\mathbf{407.8 \pm 484.0}$

actual (i.e. ground truth) count in each patch or the whole image, and $\hat{\zeta}_i$ is the estimated count. The results for both granules and images are reported as mean and standard deviation of AD.

4.3.4 Dense Crowd Density Estimation

The applicability of the proposed framework is demonstrated in the application of dense crowd density estimation on public scenes. Evaluations are conducted by benchmarking the proposed approach with state-of-the-art approaches by Idrees et al. (2013), Lempitsky and Zisserman (2010) and Rodriguez et al.(2011). These methods are among the few that is suitable for dense crowd density estimation, therefore, is used for comparison. Most existing methods (Rabaud & Belongie, 2006; Ge & Collins, 2009) require person detection, hence is more suitable for sparse crowd scenes analysis. Comparative comparison is conducted by using publicly available benchmark dataset (Idrees et al., 2013).

The comparative comparisons are presented in Table 4.1. The method by Rodriguez et al.(2011) have the highest AD per image. This is because the method relies on head detections for density estimation. For dense crowd scenes with few pixels per individual, severe occlusions and appearance variations, it is challenging to determine ones' head from another. Comparing methods by Lempitsky and Zisserman (2010) with Idrees et al. (2013), method by Idrees et al. (2013) uses three sources (i.e. head detection, SIFT and

Table 4.2: Quantitative results of the proposed approach on dense crowd density estimation using different texture features, i.e. Local Standard Deviation (LSD), Dense Scale-Invariant Feature Transform (DSIFT) and Phase congruency (PC).

Features	AD Per Patch	AD Per Image
LSD	5.9 ± 8.2	621.6 ± 679.7
LSD + DSIFT	6.7 ± 7.2	481.2 ± 523.7
LSD + DSIFT + PC	6.4 ± 6.6	407.8 ± 484.0

frequency domain analysis) whereas Lempitsky and Zisserman (2010) uses only DSIFT feature for dense crowd density estimation. Accordingly, (Idrees et al., 2013) have lower AD per image than (Lempitsky & Zisserman, 2010). This shows that to enhance density estimation in dense crowd scenes, multiple features is required to compensate and complement the insufficient of another features. By using irregular granules which conforms to the natural boundaries between crowd and background (i.e. non-crowd) regions, the proposed approach is able to extract crowd texture features essential for density estimation. This is not the case for (Idrees et al., 2013) that uses pixel-grids. Hence, the AD per patch of the proposed approach (6.4 ± 6.6) is lower than Idrees et al. (2013) (10.2 ± 18.9). Figure 4.7 shows the comparative results of dense crowd density estimation with Idrees et al. (2013).

The qualitative results of the proposed approach on dense crowd density estimation using different features are presented in Table 4.2. The first row in Table 4.2 shows the results of using Local Standard Deviation (LSD) feature only, giving AD of 5.9 ± 8.2 per patch and 621.6 ± 679.7 per image. By supplementing the proposed approach with DSIFT feature, which captures the appearance cue in dense crowd scenes, improves AD per image by 140.4. To compensate and complement the gradient-based features (LSD and DSIFT) that are sensitive to image illumination variations, PC feature that is based on phase information in frequency domain is included. This improves the AD per image to 407.8 ± 484.0. Although the mean of AD per patch increases marginally (0.5), the standard deviation reduces by 1.6.

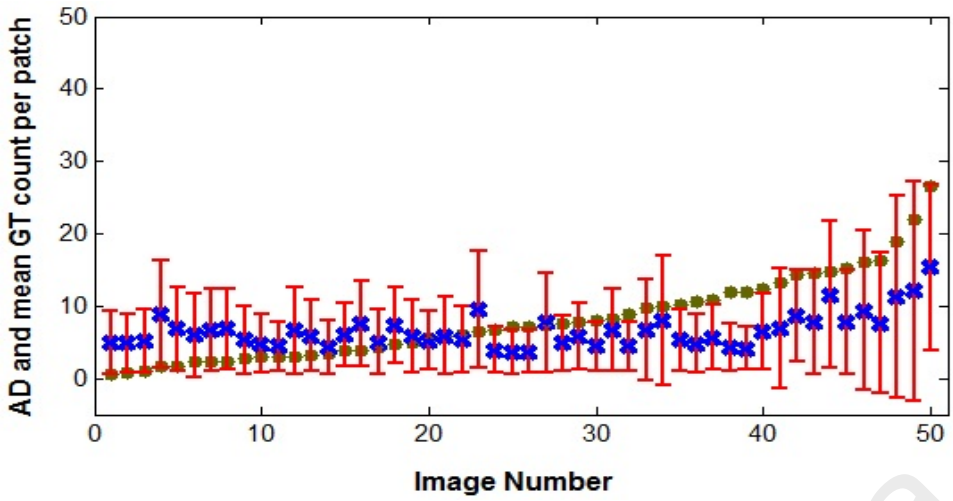


Figure 4.6: Analysis of per patch estimates in terms of absolute difference (AD). The x-axis shows image numbers sorted with respect to mean ground truth (GT) count per patch. Olive dots: GT count per patch. Blue crosses: mean of absolute difference. Red bars: standard deviation of absolute difference. Best viewed in color.

Figure 4.6 shows the AD for patches in each dense crowd image. The images are sorted with respect to the mean ground truth count per patch for ease of analysis. The mean and standard deviation of AD per patch are shown with blue crosses and red bar, respectively. The ground truth per patch for each image is shown as olive dots. As shown in Figure 4.6, the AD per patch is consistent despite the increase of ground truth count, except for the images in the range of 46 to 50. The images from the range of 1 to 45 consist of 96 – 2704 ground truth count of individual. This indicates that the proposed approach perform density estimation consistently for structure granules in this range. The reason for increasing mean and standard deviation of AD for images in the range of 46 – 50 is because these images contain the highest ground truth count, with the largest ground truth count is 4628 (i.e. a 4821% of the smallest ground truth count). Likewise, the ground truth count per patch also increases super-linearly, in contrast to the ground truth count per patch for other images. Figure 4.8 shows several dense crowd images from the dataset with their respective ground truth count and estimated count using the proposed approach.

From Figure 4.6, it is observe that there are a few images with relatively higher

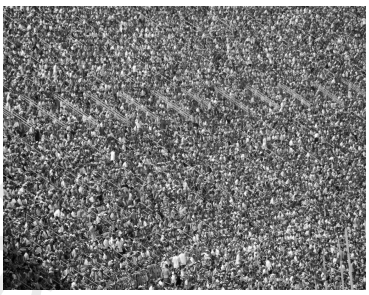
Ground Truth count: 1450
Idrees et al. (2013): 1468
Proposed: 1443

Ground Truth count: 3406
Idrees et al. (2013):1287
Estimated count: 1857

Ground Truth count: 682
Idrees et al. (2013): 653
Proposed: 653

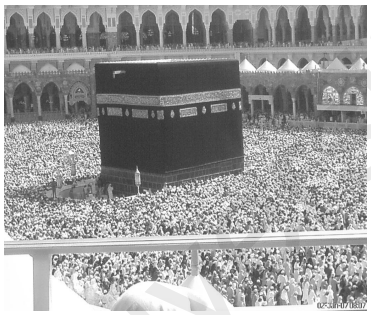
Ground Truth count: 648
Idrees et al. (2013): 640
Proposed: 694

Ground Truth count: 4628
Idrees et al. (2013): 2550
Estimated count: 1993

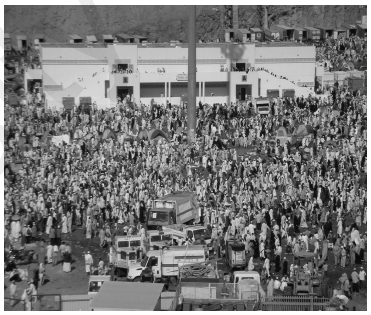


Ground Truth count: 2358
Idrees et al. (2013): 2496
Proposed: 2517

Figure 4.7: Comparative results of dense crowd density estimation with Idrees et al. (2013).



Ground Truth count: 2104
Estimated count: 2087



Ground Truth count: 1050
Estimated count: 1098

Ground Truth count: 2740
Estimated count: 1412

Ground Truth count: 2550
Estimated count: 1251

Ground Truth count: 2391
Estimated count: 1645

Ground Truth count: 967
Estimated count: 884

Figure 4.8: Several dense crowd images from the dataset with their respective ground truth count and estimated count using the proposed approach.

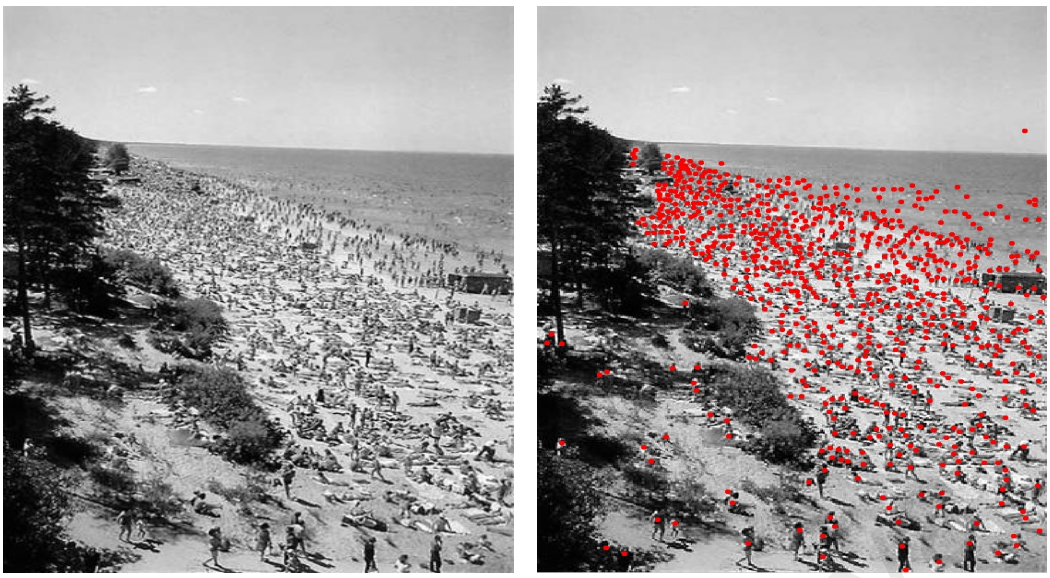


Figure 4.9: Example of low-resolution dense crowd image where it is challenging to distinguish individuals from background. Left: Dense crowd image. Right: Image with ground truth annotations (red dots). This shows that manual annotations are prone to human mistakes. Best viewed in color.

AD per patch than the overall images within the range of 1 to 45. Upon scrutinizing the results, it is observed that some of these images correspond to low resolution images where informative texture features may have been diminished. It is also challenging for human to ascertain individuals from background in the scenes (as shown in Figure 4.9). Since ground truth provided (Idrees et al., 2013) is manually annotated, it is prone to human mistake.

4.4 Summary

This chapter has explored a new approach for dense crowd density estimation by using irregular patches (i.e. granules) that conform to the natural outline between crowd and background. The proposed density estimation approach allows the granules to adapt itself to the arbitrary distribution of crowd, in which the underlying texture features characterizing crowd and background regions can be extracted. Moreover, using a set of complementing texture features is essential to compensate the insufficiencies of other features. The experimental results on public dense crowd dataset demonstrated that the use of gran-

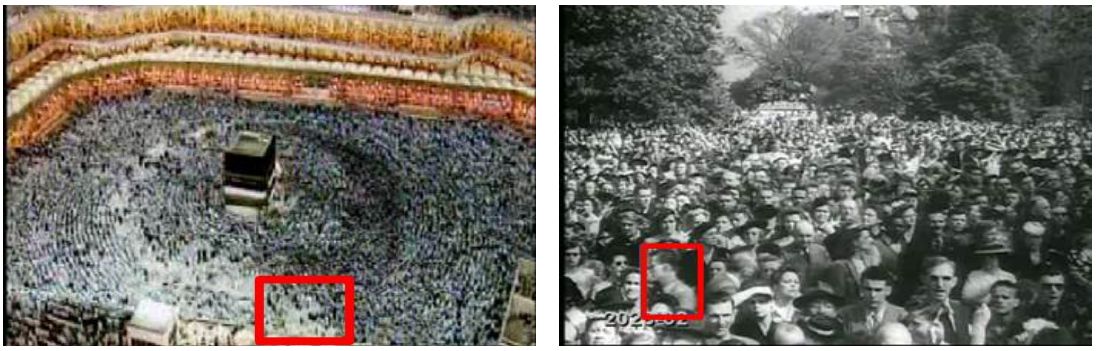
ules can improve density estimation in dense crowd scenes. Despite the importance of dense crowd density estimation research, it is acknowledged that one of the main challenges for this research is generating ground truth for evaluation. This is because manual annotation of ground truth is costly and prone to human error.

CHAPTER 5: DENSE CROWD SALIENCY DETECTION VIA GLOBAL SIMILARITY STRUCTURE

As discussed in Section 1.1, manual visual surveillance of dense crowd scenes is a formidable task, primarily due to the sheer number of individuals in scenes and excessive amount of television screens to monitor. As a result, it is not surprising for human operators that operate and monitor a set of television screens to overlook prominent events taking place within a dense crowd scene. For example, an individual that maneuvers against the dominant flow of dense crowd, as shown in Figure 5.1b. Thus, there is a dire need to computerize the processing and analysis of dense crowd to support and assist human in the detection of salient crowd regions acquiring immediate attention for a more effective surveillance.

The preceding chapters have described approaches based on GrCS to localize crowd regions and density estimation. Particularly, both chapters have demonstrated that spatial information from dense crowd scenes can serve as a useful contextual cue to facilitate dense crowd analysis. Whilst the GrCS-based method has shown superior performance in the aforementioned tasks, it is unequipped for dense crowd saliency detection. This is because it is based solely on spatial information, whereas temporal information is essential to discover and quantify saliency in dense crowd scenes.

In this chapter, a new approach is proposed for saliency detection in dense crowd scenes. The aim is to uncover the intrinsic manifold of crowd motion dynamics, which can facilitate the identification and localization of salient regions in a crowd scene. Dense crowd regions that are deemed abnormal in terms of motion flow, such as shown in Figure 5.1 (red bounding boxes) are denoted as salient. These regions are areas in dense crowd scenes with high motion dynamic. On the contrary to existing saliency detection approaches (Ali & Shah, 2007; Loy et al., 2012), low-level features extracted from crowd



(a) High motion dynamic at entry and exit points (or sources and sinks).

(b) Individual moving against the dominant dense crowd flow.

Figure 5.1: Example saliency in dense crowd scenes. These saliencies (denoted by the red bounding boxes) are area in dense crowd scenes with high motion dynamic. Best viewed in color.

motion field are transformed into a global similarity structure. The global similarity structure representation allows the discovery of the intrinsic manifold of the motion dynamics, which could not be captured by the low-level representation. Ranking is then performed on the global similarity structure to identify a set of extrema. This extrema is exploited to detect saliency in various dense crowd scenarios that exhibit crowding, local irregular motion and unique motion areas such as sources and sinks. In contrast to existing approaches (Kuettel et al., 2010; Hospedales et al., 2011; B. Zhou et al., 2012; Rodriguez, Sivic, et al., 2011), the presented manifold does not require *tracking* and *prior information or model learning* to identify interesting / salient regions in the crowded scenes. The proposed approach is thus practical for real-world dense crowd saliency detection.

The remainder of the chapter is organized as follows: Section 5.1 express the motivation of this work. This is followed by the proposed crowd saliency detection approach via global similarity structure in Section 5.2. Section 5.3 presents and discusses the experimental results. Specifically, the applicability of the proposed framework to detect saliency in dense crowd scenes is evaluated both qualitatively and quantitatively. Finally, future work are discussed and conclusion are drawn in Section 5.4.

5.1 Dense Crowd Saliency Detection

The increasing demands for security and public safety by the society has lead to an enormous growth in the deployment of closed circuit television camera (CCTV) in public spaces (Valera & Velastin, 2005; Gong, Loy, & Xiang, 2011). The recent Boston Marathon bombing, specifically, has ignited a pressing interest for automated visual analysis of dense crowd to assist the law enforcement in preventing such events from happening again. The investigation surrounding the bombing put across the fact that the incident was a missed opportunity to use technology to detect the abnormal behavior of the suspect, which leads to the tragedy (Klontz & Jain, 2013).

One must understand that at large events such as rallies and marathons, where crowds of hundreds or even thousands gather, visual monitoring is a daunting task. This is because studies (N.-H. Liu, Chiang, & Chu, 2013) have shown that the attention span of human tend to deteriorate after 20 minutes. For a monitoring task that demands prolonged cognitive attention, it is not ideal to depend solely on human operators to undertake the responsibility of identifying event in dense crowd scenes requiring immediate attention. In dense crowd scenes, any saliency or abnormal event would lead to a cascade of undesirable events because of the synergic effect of human interaction (Mehran, 2011). Consequently, major research efforts are emerging towards developing solutions to identify interesting or salient regions, which could ultimately lead to unfavorable events, as a cue to direct the attention of the security personnel.

As discussed in Section 1.1.3, the definition of interesting region in crowd has been causing much debate in the literature due to the subjective nature and complexity of the human behaviors. Some researchers consider any deviation from the ordinary observed events as anomaly, whereas others consider rare or outstanding event as interesting. Finding interesting regions in a given scene is generally accomplished by firstly learn an ac-

tivity model of the scene, followed by using the learned model to identify the anomalies (Kuettel et al., 2010; Hospedales et al., 2011; B. Zhou et al., 2012; Rodriguez, Sivic, et al., 2011).

Contrarily, the proposed approach in this chapter takes a different perspective to detect interesting regions in extremely crowded scenes. It alleviates the need of a learned model. Specifically, the motion of individuals in dense crowd scenes is assumed to follow the regular or dominant flow of a particular region due to the physical structure of the scene, and the social conventions of the crowd dynamics. With this assumption, interesting regions is considered as extrema in the underlying crowd motion dynamics in the scene. Detecting these extrema is accomplished in an unsupervised manner. In contrast to existing methods (Ali & Shah, 2007; Loy et al., 2012), which use low-level features for dense crowd motion representation, this work projects the low-level features extracted from the motion field into a global similarity structure. In this study, global similarity structure refers to the similarity / difference between every two points on a feature space, i.e. stability and phase shift. This captures the pairwise similarity of the dense crowd motion of all pixels (or particles that are spatially distributed on the image plane). Such a structure allows the discovery of intrinsic manifold of the motion dynamics. With the manifold, ranking is performed by the iterated graph Laplacian approach (D. Zhou, Weston, Gretton, Bousquet, & Schölkopf, 2004). The extrema of the rank scores are employed as an indicator of salient motion dynamics or unstable motion in the dense crowd scenes caused by crowding, sources and sinks, as well as local irregular motion.

The crowding is defined as potential clogging or bottlenecks that are typically affected by the physical structure of the environment. For example, near junctions where the crowd density builds up and thus, preventing smooth motion amongst individuals. Sources and sinks refer to regions where individuals in a crowd enter or leave the scene. Finally, local irregular motion is triggered by flow instability of individuals or a small

groups maneuvering against the dominant flow in the scene.

5.2 Proposed Dense Crowd Saliency Detection Framework

The pipeline of the proposed framework is illustrated in Figure 5.2. Given a dense crowd video sequence (Figure 5.2a), local spatiotemporal information is first extracted (Figure 5.2b). It is used to represent a set of broader definition of the crowd dynamics (Figure 5.2c and 5.2d), from which the intrinsic manifold of the motion dynamics in scenes are uncovered. Subsequently, the intrinsic manifold is exploited and used as contextual information for crowd saliency detection. In this chapter, crowd salient motions refer to conditions caused by crowding, sources and sinks, as well as local irregular motion.

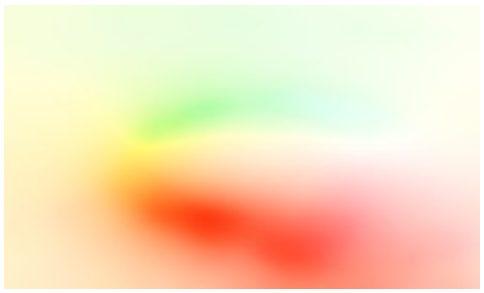
5.2.1 Crowd Motion Field

Given a dense crowd video sequence, this work commences by estimating the flow field. The proposed framework represents the crowd motion field of each frame using the optical flow. Optical flow is a velocity distribution of apparent brightness movement in an image frame, where it captures the spatio-temporal variation of pixels intensities between two frames of a video sequence (Horn & Schunck, 1981). Specifically, in a dense crowd video sequence, the velocity field at each point, $V(p) = (u_p, v_p)$ is estimated using the dense optical flow algorithm by C. Liu (2009). The velocity field at each point is calculated by its displacement between consecutive frames of a video sequence. Each pixel in a given frame is considered as a point or particle[1], $p = (x, y)$. This process is reiterated for all of the points in the frame.

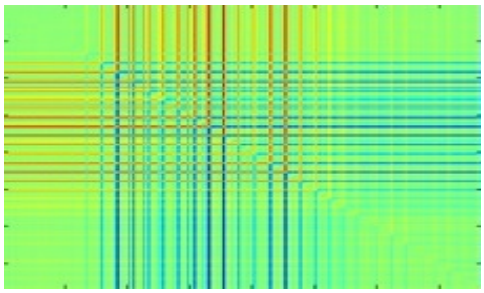
Both the horizontal and vertical velocity components, u and v, of the extracted optical flow field are then accumulated, and an averaged flow, $\overline{V}$, is calculated within an interval

[1]One could also consider a spatial block of pixels as a particle.

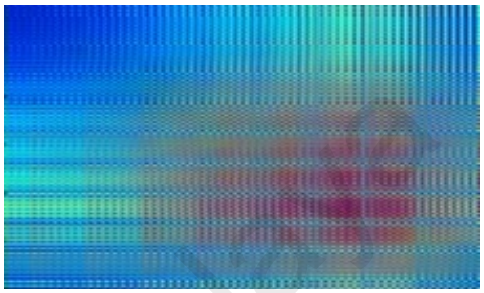
(a) Input video sequence



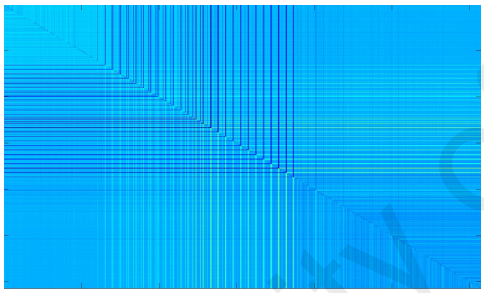
(b) Motion flow estimation



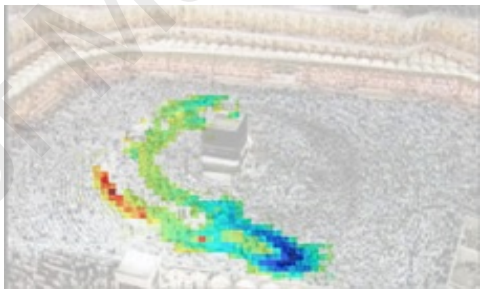
(c) Global similarity feature map: stability map



(d) Global similarity feature map: phase shift map



(e) Normalized Laplacian matrix representation



(f) The ranking results, where red and blue color indicate the extrema with interesting dynamics.

Figure 5.2: An illustration of the outputs from the key steps in crowd saliency detection. The width and height of the global similarity feature maps are the number of pixels of a video frame. Best viewed in color.

of time, comprising $|\tau|$ frames.

$$\overline{V} = \{\overline{u}, \overline{v}\} = \{\frac{1}{\tau}\sum_{t}^{t+\tau} u_p, \frac{1}{\tau}\sum_{t}^{t+\tau} v_p\} \tag{5.1}$$

Also, the mean optical flow can be denoted as $\{\overline{u}, \overline{v}\}$. Figure 5.2b shows a snapshot of the mean flow computed for the Mecca sequence.

The proposed interval-based average representation is performed to obtain smooth and consistent fields, where inconsistent velocity components (noise) are often reduced if

not removed during the averaging step.

5.2.2 Feature Representation

Using the crowd motion field, two features are extracted to represent a broader definition of the crowd dynamics for saliency detection. They are denoted as the stability and phase shift maps. These maps are the results of transformation of the low-level feature space into global similarity structure space, based on the similarity / difference between every two points on the feature space. The computation of each map will be discussed in detail in the following Section 5.2.2.1 and Section 5.2.2.2.

5.2.2.1 Stability Map

The mean optical flow field appears to be a good indicator for the dominant flow of individuals in crowd, but may not be sensitive enough to capture subtle interaction and motion flows that deviate from the norm. To this end, particle advection process is carried out under the influence of the mean optical flow field to reveal local properties of the flow. The resulting pathlines from the advection process allows quantification of the motion dynamics, which is derived later from the separation coefficients between particles.

The basic idea of particle advection is to approximate the 'transport' quantity by a set of particles as proposed by Moore et al. (2011). In this context, advection is applied to keep track of the velocity changes for each point, p along its velocity field defined by (u,v).

$$\frac{d\vec{x}_p}{dt} = u_p(t_0,t,x_0,x_p) \tag{5.2}$$

$$\frac{d\vec{y}_p}{dt} = v_p(t_0,t,y_0,y_p) \tag{5.3}$$

where (x_0,y_0) represents the initial position of point p at time t_0, while (x_p,y_p) denotes

its position at time $t_0 + t$.

Unlike the conventional optical flow representation that captures the velocity of a point in two consecutive frames, the advected flow field captures the velocity of a particle in τ consecutive frames. The trace of particles over time forms a pathline. Assuming that the initial position of p is the mean optical flow field, $\{\overline{u}, \overline{v}\}$, the problem can be formulated as an initial value problem. Cubic interpolation (Lekien & Marsden, 2005) of the neighboring flow field can be performed to compute the robust velocity of particles.

The proposed approach in this chapter adopted the Jacobian method as in Haller (2000) to measure the separation between each pathline which are seeded spatially close to a point, p, within a time instance, τ. The Jacobian is computed by the partial derivatives of $d\vec{x}_p$ and $d\vec{y}_p$, where:

$$\nabla F^t(p) = \begin{bmatrix} \frac{\partial d\vec{x}_p}{\partial x_p} & \frac{\partial d\vec{x}_p}{\partial y_p} \\ \frac{\partial d\vec{y}_p}{\partial x_p} & \frac{\partial d\vec{y}_p}{\partial y_p} \end{bmatrix} \tag{5.4}$$

According to the theory of linear stability analysis in (Seydel, 2009), the square root of the largest eigenvalue, $\lambda^t(p)$ of $F^t(p)^\top F^t(p)$ indicates the maximum offset or displacement if the particle's seeding location is shifted by one unit as it satisfies the condition that $ln\lambda^t(p) > 0$. In the context of this study, a large eigenvalue indicates that the query point is unstable, and vice versa for a small eigenvalue. In another word, a query point is regarded as stable when the velocity changes between the point and its spatially close neighboring points are minimal. Such conditions often refer to coherent motion of individuals in crowds. Conversely, unstable point demonstrates large velocity changes with respect to its spatially close neighbors.

Note that in most existing saliency detection work (Ali & Shah, 2007; Yan & Polle-

feys, 2006), unstable points are regarded as outliers or noise in video sequences and thus removed. Contrarily, in this chapter, it is believed that the key for dense crowd saliency detection is to exploit these unstable points to infer salient regions. This is due in part to the tendency of individuals to follow the dominant flow owing to the physical structure of the scene, and the social conventions of the crowd dynamics. Accordingly, any deviation in the motion dynamic of individuals from its close neighbours signifies abnormalities.

Given the eigenvalue, the stability of a point can be computed using Eq. 5.5. In practice, τ should depend on the rate of change of the flow field, with a higher rate of change of flow field resulting in smaller time scales and vice versa. In this study, τ is fixed with 50 frames at 25fps.

$$\phi^t = \frac{1}{|\tau|} \log \sqrt{\lambda^t(p)} \tag{5.5}$$

This is followed by transforming the low-level feature comprising the stability coefficient, which in this study acts as an indicator of unstable motion, into global similarity structure space. The stability map is computed by taking the difference between the stability of each point, i, with every other point, j, in the given scene:

$$s^t_{i,j} = \phi^t_i - \phi^t_j \tag{5.6}$$

where $s_{i,j}$ is the (i, j) element in the stability map denoted by $S \in \mathbb{R}^{h \times w}$, and h and w represent the height and weight of the given frame.

5.2.2.2 Phase Shift Map

In order to uncover the collective flow of a dense crowd, one of the simplest way is 'grouping' points in the average velocity field, $\overline{V}$, according to the phase similarity. Here, it is anticipated that by connecting 'grouped' points with respect to the gradual changes

of the velocity phase, can facilitate in uncovering important motion characteristic of the dense crowd.

The phase shift map is denoted by $\Theta \in \mathbb{R}^{h \times w}$. Each element $\theta^t_{i,j} \in \Theta$ is obtained as the phase difference of the mean flow vector between points:

$$\theta^t_{i,j} = \arccos \frac{\overline{V}^t_i \cdot \overline{V}^t_j}{\left\|\overline{V}^t_i\right\| \left\|\overline{V}^t_j\right\|} \tag{5.7}$$

where the phase difference, $\theta^t_{i,j}$, between two points are measured by the shortest great-circle distance, hence $\theta^t_{i,j}$ is bounded by $[0, \pi]$. The process is repeated for all the points in the average velocity field, $\overline{V}$.

The rational of projecting the velocity phase to the global similarity structure is to reveal the intrinsic relationship of each point, p, with the other points on the same video sequence.

5.2.3 Saliency Detection by Manifold Ranking

In the following, steps to detect the salient motion regions within the crowd scene by performing ranking on the intrinsic manifold (D. Zhou et al., 2004) are discussed. The intrinsic manifold is uncovered by the global similarity feature maps, i.e. the stability and phase shift maps.

For each video sequence, the set of data points $\mathscr{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$, is represented in the form of a weighted k-nearest neighbors (kNN) undirected network graph $G = \langle V, E \rangle$ to model the relations among data points. Note that each data point, $\mathbf{r} = (s^t, \theta^t)^\top$, is an integrated feature comprising the global similarity structure representation of scaled stability and phase change, where s^t and θ^t are scaled to $[0, 1]$. Each vertex, υ_i, in the graph represents a data point, $\mathbf{r}_i$. Two vertices are connected by an edge E weighted by a

pairwise affinity matrix, W_{ij}, which is defined as follows:

$$W_{ij} = \exp\left(\frac{-\text{dist}^2(\mathbf{r}_i, \mathbf{r}_j)}{\sigma_i \sigma_j}\right) \tag{5.8}$$

where $i \neq j$ and $W_{ii} = 0$ to avoid self reinforcement during the manifold ranking (D. Zhou et al., 2004). σ_i and σ_j are the local scaling parameters (Zelnik-Manor & Perona, 2004). The selection of σ_i is given as:

$$\sigma_i = \text{dist}(\mathbf{r}_i, \mathbf{r}_k) \tag{5.9}$$

where r_k is the k-th neighbor of data point r_i. The distance metric, dist, denotes the Euclidean distance. Given the affinity matrix, W_{ij}, the connected graph, G, can then be represented using the normalized Laplacian matrix, $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ (as shown in Figure 5.2e), where D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$.

Assuming that the typical and uninteresting motions dominate a dense crowd scene, thus, selecting a random set of m 'query' points, $\mathscr{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m\}$ can capture the dominant crowd behavior of the scene[2]. By performing ranking, extrema can be detected as data points with the highest and lowest rank scores, deviating from the query points. Such extrema in a video sequence suggest interesting regions caused by crowding, local irregular motion and sources and sinks.

To detect the extrema, each query are labelled successively with a positive label +1. Its label is then propagated to all other unlabeled instances, $\{\mathbf{r}_i\}$, of which their initial labels are assigned as 0. More precisely, a rank score vector is computed for each query

[2]The selection of those random points can be repeated to generate more queries, accordingly. In this study, m has been set to 100. Evaluation with varying query points generated consistent rank score.

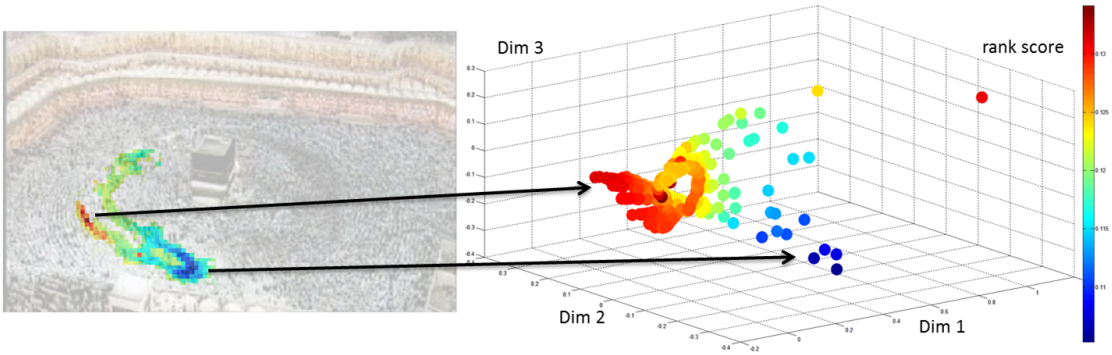


Figure 5.3: Three-dimensional embedding of the global similarity structure obtained using multi-dimensional scaling. The color of each point represents the ranking score, where the extrema correspond to salient regions. Best viewed in color.

$\mathbf{q}_i$, individually, denoted as $\mathbf{c}_i = (c_i^1, ..., c_i^n)^\top$, via the normalized Laplacian matrix, L, using the close form equation:

$$\mathbf{c}_i = (I - \alpha L)^{-1} y \tag{5.10}$$

where I is an identity matrix and α is a scaling parameter in the range of $[0,1]$. The vector y is the initial label assignment of data points, which is given as $y = (y_1, ..., y_n)^\top$, in which $y_i = +1$ if $\mathbf{r}_i = \mathbf{q}_i$, and $y_i = 0$ otherwise. Note that $\mathbf{q}_j$ where $j \neq i$ has initial label assigned as 0 too. The same ranking process is repeated for all query points $\mathcal{Q}$. The final rank score vector, $\mathbf{C}$, is the average of m rank score vectors, i.e. $\mathbf{C} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{c}_i$. Extrema are data points with the highest and the lowest rank scores in $\mathbf{C}$, as illustrated in Figure 5.3.

5.3 Experiments

In the following sections, the dataset used in the experiments, experimental setup and saliency detection results on dense crowd scenes are discussed.

5.3.1 Dataset

Evaluation on the proposed saliency detection framework are conducted on 22 benchmark dataset of public dense crowd scenes obtained from Rodriguez, Sivic, et al. (2011), Loy et al. (2012) and Solmaz et al. (2012). These dataset are diverse, where it consists of dense crowd scenes in various scenarios, such as pilgrimage, marathon, station, rallies and stadium. In addition, the sequences have different field of views, resolutions, and exhibit a multitude of motion behaviors that cover both the obvious and subtle instabilities. To evaluate the efficiency of the proposed framework in saliency detection, the ground truth of dense crowd saliencies is manually annotated by exhaustive frame-wise examination on the entire dataset. Examples of dense crowd scenes from the dataset and the respective ground truth annotation (i.e. blue bounding boxes) are shown in Figure 5.4.

5.3.2 Experiment Settings

The proposed framework was developed in Matlab-r2013a environment. Experimental evaluations are performed on a computer with 64-bit Microsoft Windows 7 operating system with 3.40 GHz Intel(R) Core i7-3770 processor. In all these experiments, the time instance, τ is set to be 50 frames at 25 frames per second.

5.3.3 Qualitative Analysis

The proposed framework is assessed in the application of dense crowd saliency detection. Evaluations are conducted by benchmarking the proposed framework with conventional approaches by Loy et al. (2012), Solmaz et al. (2012) and Ali and Shah (2007). Each evaluation is compared against the benchmark dataset used in each respective approach.

The qualitative evaluation is divided into two assessments: instability detection and local irregular motion detection.

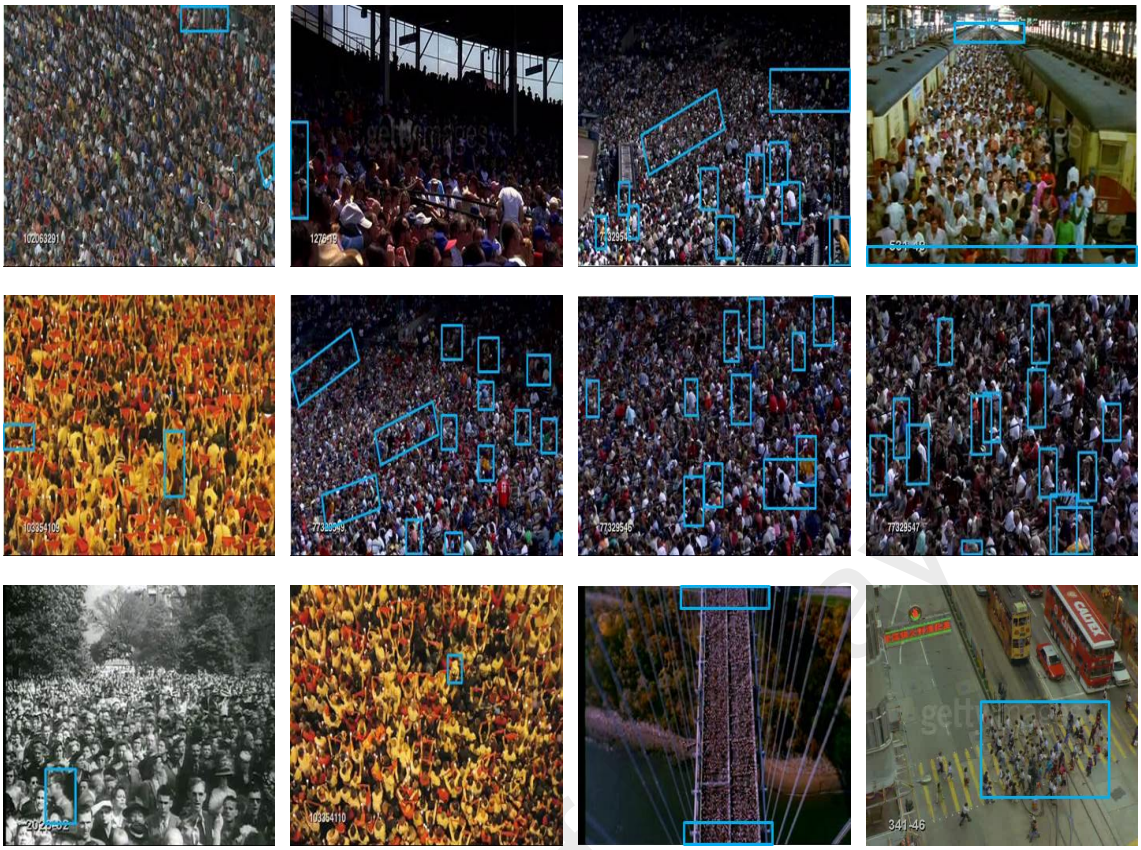


Figure 5.4: Example dense crowd sequences from the dataset on which experiments were performed with the corresponding ground truth annotations (i.e. blue bounding box). The sequences in the dataset consist of dense crowd in various scenarios, such as parades, concerts and rallies. The saliencies (annotated by the blue bounding box) are areas in dense crowd with high motion dynamic. Best viewed in color.

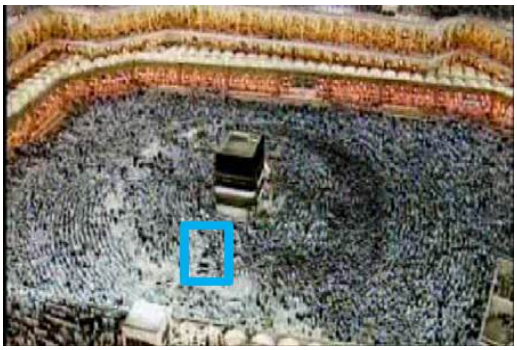
5.3.3.1 Instability Detection

A set of two sequences comprising a pilgrimage and marathon scenes were used to test the capability of the proposed framework in detecting instability. Manual annotation of saliencies in such dense crowd scenes is nontrivial given the high level of activities and interactions among individuals. Moreover, as pointed out by Lim (2014), the saliencies / anomalies which lead to various dense crowd disasters are ambiguous in nature. Therefore, following the studies by Loy et al. (2012) and Ali and Shah (2007), synthetic instabilities are inserted into the two original video sequences to simulate unstable region. The synthetic instabilities are as enclosed in the blue bounding box shown in Figure 5.5a and red bounding box in Figure 5.6a. These instabilities are created by randomly posi-

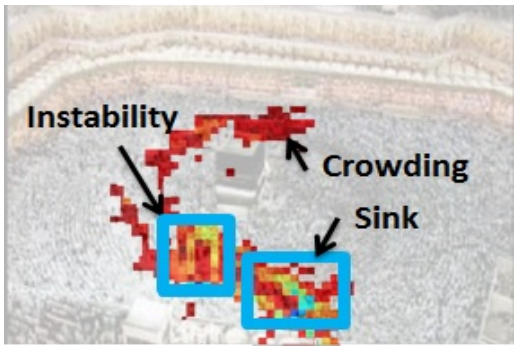
tioning a bounding box within the dense crowd motion flow, and subsequently flipping and rotating it to alter the flow at that location (Ali & Shah, 2007). The spatial positions of these instabilities are noted to serve as the ground truth for experimental evaluations.

As shown in Figure 5.5 and Figure 5.6, the proposed approach is able to accurately identify the regions injected with synthetic instability. This is analogous to approaches by Loy et al. (2012) and Ali and Shah (2007). Additionally, the proposed approach is able to identify other regions that exhibit high motion dynamics as highlighted by the colored regions. On closer scrutiny, it is observed that these areas correspond to the exit and turning point around the Kaaba in Figure 5.5. There is potential slowdown in the pace of individuals and shift of walking direction due to the structure of the environment and change in the intensity of physical constrains between individuals in that region. Similarly, the proposed approach is able to detect the sink region in the marathon sequence in Figure 5.6, where the crowd exit from the field of view. The fact that there are only a few pixels per individual and high level of interaction among them makes such saliencies challenging even for human visual perception to notice. Nonetheless, the results demonstrate the effectiveness of the global similarity structure in capturing the intrinsic structure of the crowd motion for saliency detection.

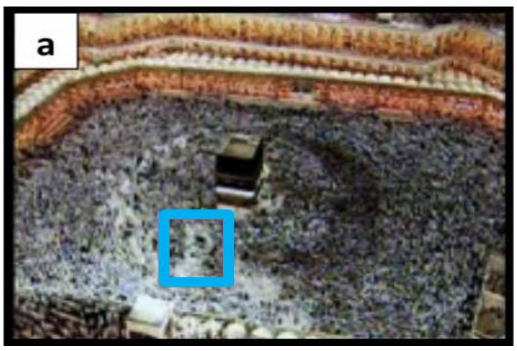
To further evaluate the proposed method in dealing with inconsistent and subtle crowd motion, the proposed approach and the benchmark approaches by Loy et al. (2012) and Ali and Shah (2007) are tested on the original sequences of pilgrimage and marathon, without any synthetic instability. The results in Figure 5.7 and Figure 5.8 show that approaches by Loy et al. (2012) and Ali and Shah (2007) do not have any detection for these sequences. On the contrary, the proposed approach is capable of detecting the sink region, as well as the potential overcrowding regions along the bridge's edge. Note that the results herein are consistent with the sequences with synthetic instability where the proposed approach detects similar interesting regions. The results, again, show that subtle



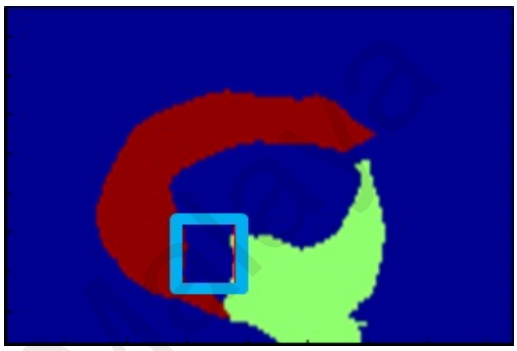
(a) Original image (with synthetic instability)



(b) Proposed approach

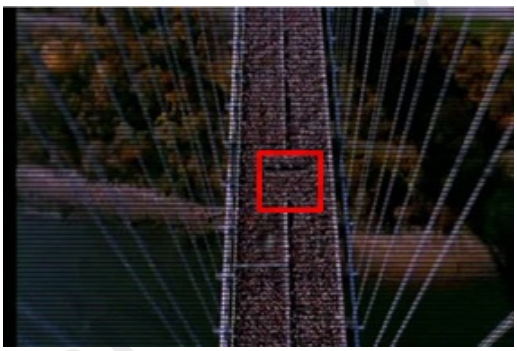


(c) Loy et al. (2012)

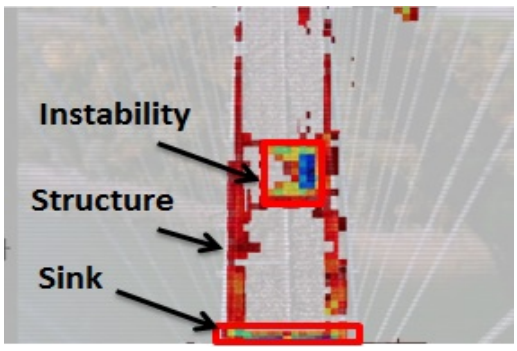


(d) Ali and Shah (2007)

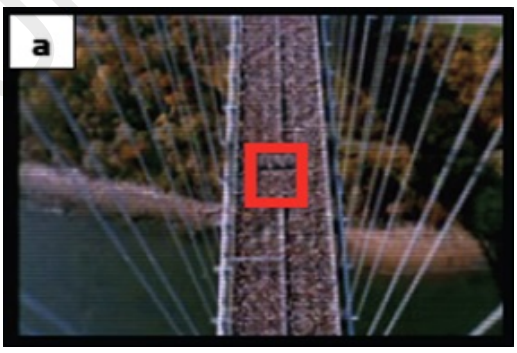
Figure 5.5: Comparisons on the corrupted pilgrimage sequence, where synthetic instability was added to simulate unstable motion. Best viewed in color.



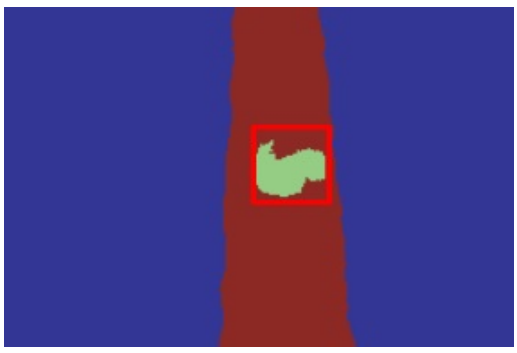
(a) Original image (with synthetic instability)



(b) Proposed approach

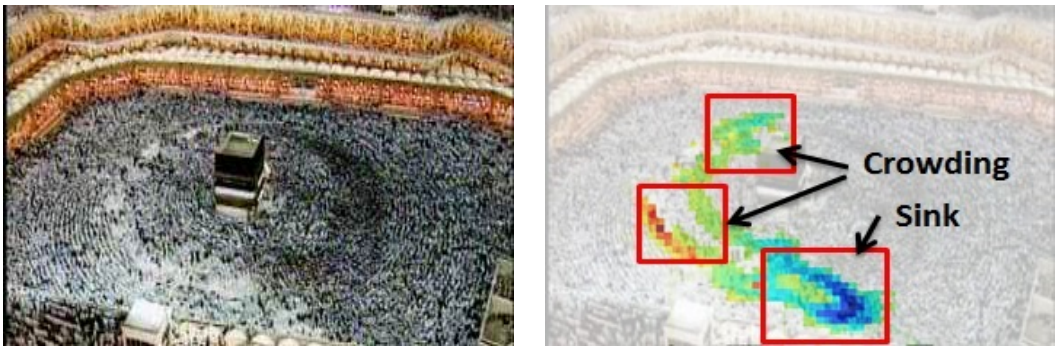


(c) Loy et al. (2012)



(d) Ali and Shah (2007)

Figure 5.6: Comparisons on the corrupted marathon sequence, where synthetic instability was added to simulate unstable motion. Best viewed in color.



(a) Original image **(b)** Proposed approach

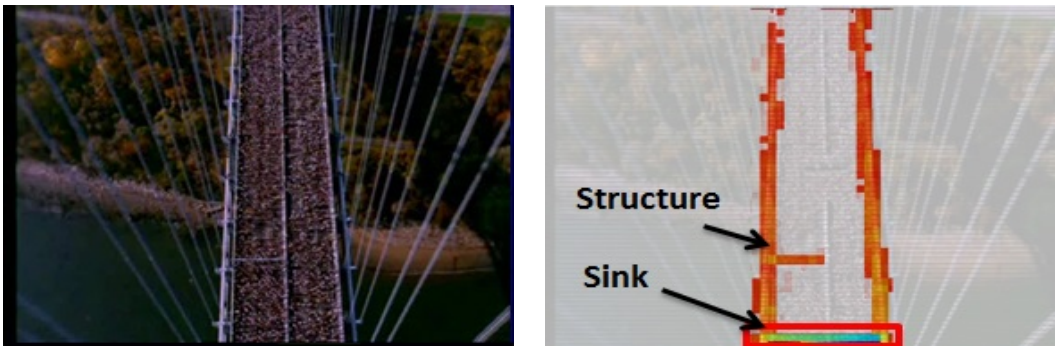
(c) Loy et al. (2012) **(d)** Ali and Shah (2007)

Figure 5.7: Comparisons on the original pilgrimage sequence (without synthetic instability). Best viewed in color.

motion can be more effectively discovered by employing the global similarity structure of the crowd motion rather than using the low-level flow field (Loy et al., 2012; Ali & Shah, 2007).

5.3.3.2 Local Irregular Motion Detection

Another comparative comparison is performed between the proposed approach and Solmaz et al. (2012) using the sequence obtained from an underground station as depicted in Figure 5.9. This sequence contains obvious source and sink regions, which are detected as bottleneck and fountainhead in (Solmaz et al., 2012). The results demonstrate that the proposed approach is able to detect similar regions as in (Solmaz et al., 2012), with the addition of another source region at the bottom right of the scene, which is not detected by Solmaz et al. (2012). Furthermore, the proposed approach is able to detect the irregular motion of someone walking into the scene from the bottom left corner of the scene. This



(a) Original image **(b)** Proposed approach

(c) Loy et al. (2012) **(d)** Ali and Shah (2007)

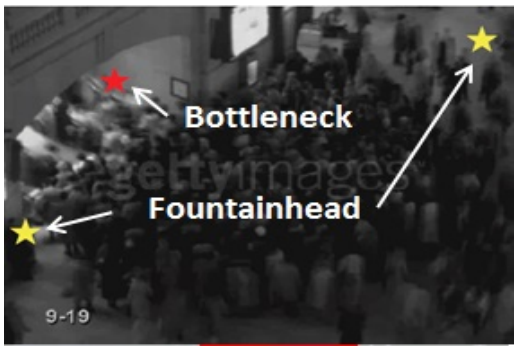
Figure 5.8: Comparisons on the original marathon sequence (without synthetic instability). Best viewed in color.

is not the case in (Solmaz et al., 2012), where their detection does not highlight accurately the location of the triggering event. It is worth pointing out that while the proposed approach is able to detect salient/interesting motion dynamics, these saliencies are not characterized into different categories.

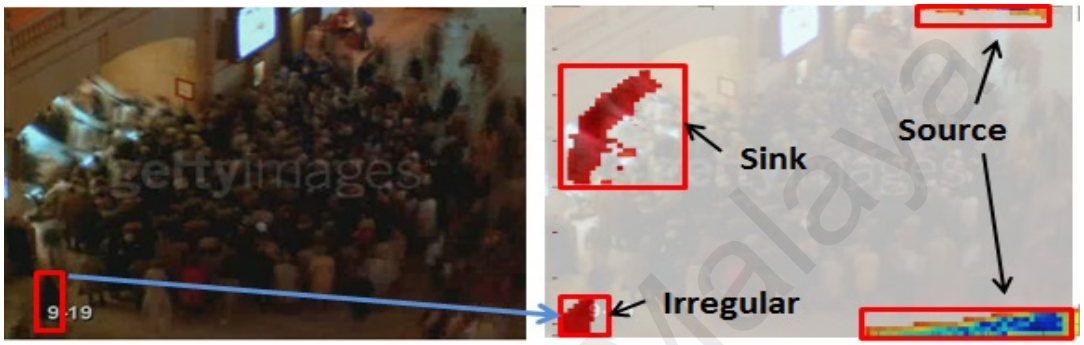
Further evaluations are conducted to test the proposed approach on sequences with local irregular motion caused by individuals moving against the dominant crowd flow such as that shown in Fig. 5.10. This scenario is to mimic the Boston Marathon Person Finder page[3] launched by Google, which aims to identify individuals that seem suspicious. Through the proposed global similarity structure of the dense crowd motion, the proposed approach is able to detects such anomaly, as illustrated in Figure 5.10.

[3]Person Finder: Boston Marathon Explosions: https://google.org/personfinder/2013-boston-explosions/

(a) Original image



(b) Solmaz et al. (2012)



(c) Proposed approach

Figure 5.9: Comparison with the state-of-the-art method by Solmaz et al. (2012) on the station sequence. Best viewed in color.

(a) An individual walking across the scenes, when the rest of the crowd is seated.

(b) An individual maneuvering through a dense crowd.

Figure 5.10: Example detections on local irregular motion. The ground truth is enclosed in the white bounding box in the first two columns. Saliency detection output from the proposed approach is highlighted in the blue bounding box on the right column. (a) Proposed approach detects an individual walking across the scene, while the rest of the crowd is seated. (b) Proposed approach detects an individual maneuvering through a dense crowd. Best viewed in color.

Table 5.1: Summary of the dense crowd saliency detection results.

Motion Category	Total # of Labelled Region	# of Detection	# of Missed Detection	# of False Detection
Crowding	13	12	1	0
Sources & Sinks	19	14	5	0
Local Irregularity	43	47	2	6

5.3.4 Quantitative Analysis

As research on dense crowd saliency detection expand, public dataset start to gain importance to meet the research requirement. Nonetheless, the comparative comparisons and benchmark datasets developed are characteristically of its own. Most of the related studies (Loy et al., 2012; Ali & Shah, 2007), merely provide qualitative results and the implementations are not shared publicly; leading to difficulties in performing a comprehensive evaluation quantitatively. As such, for quantitative evaluation of the proposed saliency detection approach, ground truth was obtained by exhaustive frame-wise examination. The regions with interesting motion dynamics are determined as per video basis where the *F-measure* according to the score measurement of the well-known PASCAL challenge (Everingham et al., 2010) is applied. That is, if the detected region overlaps the ground truth region by more than 50 %, then the detection is considered as the correct salient region. The propose approach are compared against the generated ground truth for all the sequences on the public datasets.

For clarity, the detection results are presented according to the different interesting motion categories, i.e. crowding, sources and sinks and local irregular motion, as shown in Table 5.1. In general, the proposed approach is able to detect saliencies in dense crowd scenes with only several false detections that are due to ambiguous local motion, i.e. random hand waving motion in a crowded scene. Figure 5.11 shows additional output of saliency detection on public dataset.

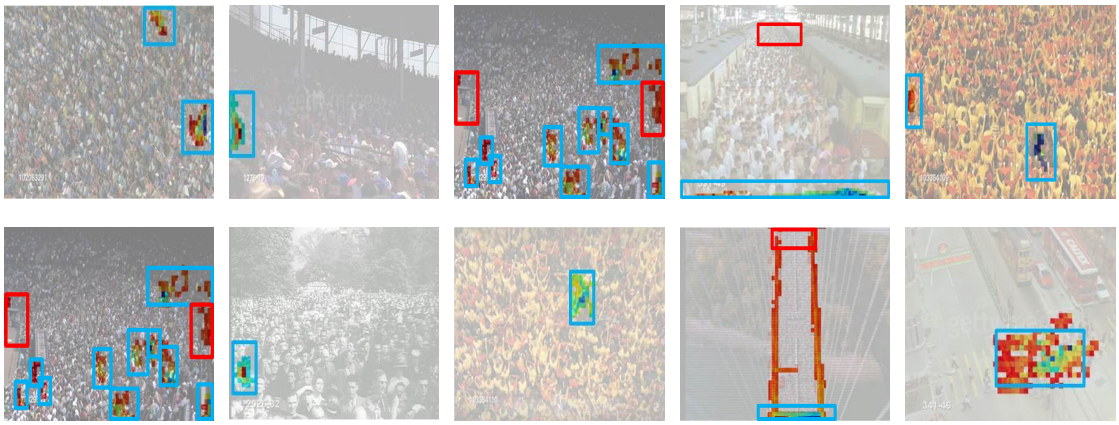


Figure 5.11: Additional results of saliency detection on public dataset. Blue bounding box: true positive (accurate detection). Red bounding box: false positive and false negative (inaccurate detection). Best viewed in color.

5.4 Summary

This chapter has presented a new approach for saliency detection in dense crowd scenes by preserving pairwise similarity of motion features. In particular, low-level motion features, i.e. stability and phase shift, extracted from dense crowd motion field are transformed into global similarity structure, based on the similarity / difference between every two points on the feature space. Experimental results on public dataset have shown that the method is effective in detecting sources and sinks, crowding, and local irregular motions from various dense crowd scenes. Importantly, the proposed approach does not require tracking of individuals in crowd. Consequently, as demonstrated through experiments in this chapter, it can be applied to dense crowd scenes where inter-occlusions due to the sheer number of individuals are apparent. Moreover, no prior information or model learning is required to identify interesting / salient regions in the scene, since extrema in the intrinsic manifold of motion dynamics is used as an indicator of saliency.

Though the global similarity structure representation allows the discovery of the intrinsic manifold of the motion dynamics, the basis of the proposed approach is optical flow. Thus, it is confined by the drawback of optical flow which assumes brightness constancy. Specifically, the underlying assumption is such that the intensities of objects

remains fixed from one frame to the next. This assumption rarely holds true for real-world scenes. Future investigation includes identifying low-level features that are robust towards characterising motion in dense crowd scenes. In addition, it is acknowledged that the subjectivity of saliency in dense crowd scenes poses considerable challenges for ground truth annotation. This in turn led to difficulty to establish a benchmark dataset for comparative comparison, both qualitatively and quantitatively. Nonetheless, the saliencies detected using the proposed approach can be a mean to support human in visual surveillance of dense crowd scenes.

CHAPTER 6: CONCLUSION AND FUTURE WORK

This thesis has been devoted toward visual analysis of dense crowds using computer vision techniques. Specifically, the thesis addresses the three researches associated with dense crowds: (1) localization of crowd segments in public scenes, (2) density estimation in dense crowd scenes, and (3) detection of unusual events in crowded public scenes, to assist human in improving dense crowd safety and security.

These research problems are nontrivial owing to the sheer number of individuals in scenes, which lead to severe occlusions among individuals. Perspective distortion due to camera orientation and position, as well as visual ambiguities further complicates the problems, resulting in appearance variations of crowds. Moreover, the unpredictability of unusual events and the elusive representation of abnormality in dense crowd scenes remains an issue. As concluded in Chapter 2, the overview of the available literature suggest that there is still a considerable scope to improve visual analysis of dense crowds.

6.1 Dense Crowd Segmentation

This thesis has presented an alternative approach to localize crowd segments in public dense crowd scenes in Chapter 3. Specifically, a new approach has been proposed using the concept and principles of granular computing (GrC) to simplify dense crowd scenes into structurally meaningful atomic regions (i.e. granules) for dense crowd segmentation. These granules are utilized to obviate the difficulty of segregating individuals in dense crowd due to context variations of crowd, by enabling inference of crowd and background regions based on local structures. The correlation among image granules at different levels of granularity is exploited to alleviate the difficulty of defining the natural boundaries between crowd and background (i.e. non-crowd) regions. These granules conformed to the boundaries of crowd segments, thus have the advantage of being scene-independent.

The proposed approach has shown superior performance as compared to existing techniques in segmenting dense crowd regions on public and synthetic dense crowd scenes.

With regard to granular-based dense crowd segmentation, there are several possible extensions. Despite the fact that the granules is effective in outlining boundaries between crowd and background regions, the basis of granules are texture features on spatial domain. Thus, granulated views of different granularities are susceptible to illuminations changes. Future investigation includes identifying features that are more robust toward characterizing poor illuminated crowd scenes. Also, although the image-based approach proposed in Chapter 3 allow separation between texture features of crowd and background, video-based approach has the advantage of temporal information. The video-based crowd segmentation approach, generally, captures motions of crowd throughout a sequence of images (Ali & Shah, 2007). The temporal information from video-based method can, therefore, be utilized to complement the texture features used in the proposed approach to enhance dense crowd segmentation.

6.2 Density Estimation

Extending from Chapter 3, this thesis demonstrated the importance of using granules for density estimation in Chapter 4. Particularly, unlike existing techniques that uses pixel-grid (Idrees et al., 2013; Marana et al., 1998), the proposed dense crowd density estimation approach used granules that conform to natural outline of crowds. These granules serve as meaningful primitive regions to extract features for density estimation. The features extracted are exploited to establish a direct mapping to the actual people count. Experimental results have shown that the proposed strategy outperformed existing pixel-grid based approach in estimating density of dense crowds.

In Chapter 2, the available density estimation literature suggests that most of the research has been focused on scenes containing low density of people. Hence, research on

dense crowd density estimation lacks standard dataset and performance framework for benchmarking purposes. Thus, amongst the future work in this aspect is to collect a more comprehensive dataset for benchmarking within the research community. However, it is acknowledged that generating ground truth for evaluation involves manual annotations. The annotation process can be costly and prone to human error. Another aspect for further analysis is to include information of crowd motion dynamics as features to reduce ambiguities of texture features while improving the accuracy of density estimation.

6.3 Saliency Detection

Apart from dense crowd segmentation and density estimation, this thesis also presents a new approach for saliency detection in dense crowd scenes. The proposed approach described in Chapter 5 transforms low-level features, i.e. stability and phase shift, extracted from crowd motion field into a global similarity structure. This is to uncover intrinsic manifold of crowd motion dynamic to facilitate the localization of salient regions. By performing ranking on the global similarity structure, the experimental results have shown that it can detect sources and sinks, crowding, and local irregular motions from various dense crowd scenes. Importantly, this is achieved without the need of person tracking. As demonstrated through experiments, the proposed approach can thus be applied on dense crowd scenes where occlusions among individual in dense crowd is prominent. In addition, the proposed approach alleviate the need of prior information or model learning to identify salient regions in scenes, since extrema in the intrinsic manifold of motion dynamics from ranking is used as an indicator of saliency.

Whilst this work has demonstrated the effectiveness of using crowd motion dynamic for saliency detection, the basis of the motion dynamic optical flow. It is, thus, confined by the drawback of optical flow which assumes that the intensities of objects remains fixed from one frame to another. Future analysis includes identifying low-level features

that are robust towards characterizing motions in dense crowd scenes. As research on saliency detection in dense crowd scenes expand, public dataset start to gain importance to meet the research requirement. However, the comparative comparison and benchmark datasets developed are characteristically of its own. This is predominantly due to the subjectivity of the definition of saliency in dense crowd scenes, leading to difficulty to summarize the evaluation protocol and performance comparison in this research field. Thus, future extension includes preparing a comprehensive public dataset and a common platform for performance comparison.

Current proposed approach focuses on analyzing regions in crowd with high motion dynamic to infer saliency. It would provide richer information of the scenes if an algorithm could localize and analyze regions of sudden non-moving crowd as well. This counter-intuitive approach is based on the notion that individuals or group that stop abruptly are worthy of attention (Yi, Li, & Wang, 2015).

6.4 Summary

Visual analysis of dense crowds is nontrivial owing to the sheer number of individuals and interactions among individuals in scenes. However, with the steady worldwide population growth and continuing urbanization, research on visual analysis of dense crowds in the field of computer vision has prospered, with the aim to assist human in visual analysis task. At present, this field of research is still in its infancy stage and requires ongoing research efforts. In an effort toward the development of dense crowd analysis research, this thesis has presented several approaches that obviate the difficulty of segregating each individual in analyzing dense crowds. Specifically, this thesis focuses on dense crowd segmentation, density estimation and saliency detection, to achieve a collective analysis of dense crowds.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2010). Slic superpixels. *Ecole Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep*, *2*, 3.

Aggarwal, C. C. (2004). A human-computer interactive method for projected clustering. *IEEE Transactions on Knowledge and Data Engineering*, *16*(4), 448–460.

Aggarwal, J. K., & Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, *73*(3), 428–440.

Ahmed, Q. A., Arabi, Y. M., & Memish, Z. A. (2006). Health risks at the hajj. *The Lancet*, *367*(9515), 1008–1015.

Ali, S. (2008). *Taming crowded visual scenes* (PhD thesis). University of Central Florida, Florida, United States of America. (AAI3377797)

Ali, S., Nishino, K., Manocha, D., & Shah, M. (2013). Modeling, simulation and visual analysis of crowds: A multidisciplinary perspective. In *Modeling, simulation and visual analysis of crowds* (Vol. 11, pp. 1–19). New York, United States of America: Springer New York.

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

Ali, S., & Shah, M. (2008). Floor fields for tracking in high density crowd scenes. In *Proceeding of European Conference on Computer Vision* (pp. 1–14).

Allain, P., Courty, N., & Corpetti, T. (2012). AGORASET: a dataset for crowd video analysis. In *ICPR International Workshop on Pattern Recognition and Crowd Analysis* (pp. 1–6).

Andrade, E. L., Blunsden, S., & Fisher, R. B. (2006). Modelling crowd scenes for event detection. In *International Conference on Pattern Recognition* (pp. 175–178).

Arandjelovic, O. (2008). Crowd detection from still images. In *Proceedings of the British Machine Vision Association Conference* (pp. 1–10).

Bargiela, A., Pedrycz, W., & Hirota, K. (2004). Granular prototyping in fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *12*(5), 697–709.

Bokhary, K. (1993). *The Lan Kwai Fong disaster on January 1, 1993: Final report*. Hong Kong: Government Printer.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Cao, L., Zhang, X., Ren, W., & Huang, K. (2015). Large scale crowd analysis based on convolutional neural network. *Pattern Recognition*, *48*(10), 3016 – 3024.

Chan, A. B. (2008). *Beyond dynamic textures: A family of stochastic dynamical models for video with applications to computer vision* (PhD thesis). University Of California, San Diego.

Chan, A. B., & Dong, D. (2011). Generalized gaussian process models. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2681–2688).

Chan, A. B., Liang, Z.-S. J., & Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–7).

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

Chen, D.-Y., & Huang, P.-C. (2011). Motion-based unusual event detection in human crowds. *Journal of Visual Communication and Image Representation*, *22*(2), 178–186.

Chen, J., Shan, S., Zhao, G., Chen, X., Gao, W., & Pietikäinen, M. (2008). A robust descriptor based on weber's law. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–7).

Chen, K., Gong, S., Xiang, T., & Loy, C. C. (2013). Cumulative attribute space for age and crowd density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2467–2474).

Chen, K., Loy, C. C., Gong, S., & Xiang, T. (2012). Feature mining for localised crowd counting. In *Proceedings of the British Machine Vision Association Conference* (Vol. 1, pp. 21.1–21.11).

Chiappino, S., Morerio, P., Marcenaro, L., & Regazzoni, C. S. (2014). Bio-inspired relevant interaction modelling in cognitive crowd management. *Journal of Ambient Intelligence and Humanized Computing*, *6*(2), 171–192.

Chongjing, W., Xu, Z., Yi, Z., & Yuncai, L. (2013). Analyzing motion patterns in crowded scenes via automatic tracklets clustering. *Communications, China*, *10*(4), 144–154.

Courty, N., Allain, P., Creusot, C., & Corpetti, T. (2014). Using the agoraset dataset: Assessing for the quality of crowd video analysis methods. *Pattern Recognition Letters*, *44*, 161–170.

Darby, P., Johnes, M., & Mellor, G. (Eds.). (2005). *Soccer and disaster: International perspectives*. London, England: Routledge.

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

Dee, H. M., & Hogg, D. (2004). Detecting inexplicable behaviour. In *Proceedings of the British Machine Vision Association Conference* (pp. 1–10).

Dehghan, A., Idrees, H., Zamir, A. R., & Shah, M. (2014). Automatic detection and tracking of pedestrians in videos with various crowd densities. In *Pedestrian and evacuation dynamics* (pp. 3–19). Springer.

Deneubourg, J.-L., Pasteels, J. M., & Verhaeghe, J.-C. (1983). Probabilistic behaviour in ants: a strategy of errors? *Journal of Theoretical Biology*, *105*(2), 259–271.

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms* (Tech. Rep.). Technical report, Department of Computer Science, Oregon State University.

Digital Design & Imaging Service Inc. (2015). *Crowd counting and analysis services.* Retrieved from `http://airphotoslive.com/portfolio/crowd-counting-and-analysis-services/`

Dong, L., Parameswaran, V., Ramesh, V., & Zoghlami, I. (2007). Fast crowd segmentation using shape indexing. In *IEEE International Conference on Computer Vision* (pp. 1–8).

Dubos, R. (1974). The social environment. In C. M. Loo (Ed.), *Crowding and behavior*

(pp. 55–60). Ardent Media.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fagette, A., Courty, N., Racoceanu, D., & Dufour, J.-Y. (2014). Unsupervised dense crowd detection by multiscale texture analysis. *Pattern Recognition Letters*, *44*, 126–133.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8).

Ferryman, J., & Ellis, A. (2010). PETS2010: Dataset and challenge. In *IEEE Conference on Advanced Video and Signal Based Surveillance* (pp. 143–150).

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Gad-el Hak, M. (Ed.). (2008). *Large-scale disasters: Prediction, control, and mitigation*. Cambridge, England: Cambridge University Press.

Ge, W., & Collins, R. T. (2009). Marked point processes for crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2913–2920).

Ge, W., & Collins, R. T. (2010). Crowd detection with a multiview sampler. In *Proceeding of European Conference on Computer Vision* (pp. 324–337). Springer.

Ghidoni, S., Cielniak, G., & Menegatti, E. (2013). Texture-based crowd detection and localisation. In *Intelligent Autonomous Systems* (pp. 725–736). Springer.

Gong, S., Loy, C. C., & Xiang, T. (2011). Security and surveillance. In *Visual Analysis of Humans* (pp. 455–472). Springer.

Haller, G. (2000). Finding finite-time invariant manifolds in two-dimensional velocity fields. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *10*(1), 99–108.

Helbing, D., Brockmann, D., Chadefaux, T., Donnay, K., Blanke, U., Woolley-Meza, O.,

... Perc, M. (2014). Saving human lives: What complexity science and information systems can contribute. *Journal of Statistical Physics*, *158*(3), 735–781.

Helbing, D., Farkas, I. J., Molnar, P., & Vicsek, T. (2002). Simulation of pedestrian crowds in normal and evacuation situations. *Pedestrian and Evacuation Dynamics*, *21*(2), 21–58.

Helbing, D., Johansson, A., & Al-Abideen, H. Z. (2007). Dynamics of crowd disasters: An empirical study. *Physical Review E*, *75*(4), 046109-1–046109-7.

Helbing, D., Molnar, P., Farkas, I. J., & Bolay, K. (2001). Self-organizing pedestrian movement. *Environment and Planning B*, *28*(3), 361–384.

Helbing, D., & Mukerji, P. (2012). Crowd disasters as systemic failures: Analysis of the love parade disaster. *EPJ Data Science*, *1*(1), 1–40.

Hendee, W. R., & Wells, P. N. (1997). *The perception of visual information*. United States of America: Springer Science & Business Media.

Heppner, F., & Grenander, U. (1990). A stochastic nonlinear model for coordinated bird flocks. *American Association for the Advancement of Science*, *89*, 233–238.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoo, W. L., Kim, T.-K., Pei, Y., & Chan, C. S. (2014). Enhanced random forest with image/patch-level learning for image understanding. In *International Conference on Pattern Recognition* (pp. 3434–3439).

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. In *1981 Technical symposium east* (pp. 319–331).

Hospedales, T. M., Li, J., Gong, S., & Xiang, T. (2011). Identifying rare and subtle behaviors: A weakly supervised joint topic model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2451–2464.

Hou, Y.-L., & Pang, G. (2013). Multicue-based crowd segmentation using appearance and motion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *43*(2), 356–369.

Hsu, E. B., & Burkle, F. M. J. (2012). Cambodian Bon Om Touk stampede highlights preventable tragedy. *Prehospital and Disaster Medicine*, *27*(5), 481–482.

Hu, M., Ali, S., & Shah, M. (2008). Learning motion patterns in crowded scenes using motion flow field. In *Proceedings of the International Conference on Pattern Recognition* (pp. 8–11).

Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., & Maybank, S. (2006). A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(9), 1450–1464.

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

Idrees, H., Soomro, K., & Shah, M. (2015). Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(10), 1986–1998.

Idrees, H., Warner, N., & Shah, M. (2014). Tracking in dense crowds using prominence and neighborhood motion concurrence. *Image and Vision Computing*, *32*(1), 14–26.

Ihaddadene, N., & Djeraba, C. (2008). Real-time crowd motion analysis. In *International Conference on Pattern Recognition* (pp. 1–4).

Junior, J. S. J., Musse, S., & Jung, C. (2010). Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, *5*(27), 66–77.

Kaiman, J. (2015). Shanghai: dozens killed and injured in stampede at new year celebrations. *The Guardian*. Retrieved from `http://www.theguardian.com/world/2014/dec/31/shanghai-35-people-killed-42-injured-new-year-crush`

Kang, K., & Wang, X. (2014). Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*.

Klontz, J. C., & Jain, A. K. (2013). A case study of automated face recognition: The boston marathon bombings suspects. *Computer*, *46*(11), 91–94.

Kong, D., Gray, D., & Tao, H. (2006). A viewpoint invariant approach for crowd counting. In *International Conference on Pattern Recognition* (Vol. 3, pp. 1187–1190).

Kovesi, P. (1999). Image features from phase congruency. *Videre: Journal of Computer Vision Research*, *1*(3), 1–26.

Kovesi, P. (2000). Phase congruency: A low-level image invariant. *Psychological Research*, *64*(2), 136–148.

Krantz, S. G. (2004). *Real analysis and foundations, second edition*. United States of America: Chapman and Hall.

Kratz, L., & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1446–1453).

Kuettel, D., Breitenstein, M. D., Van Gool, L., & Ferrari, V. (2010). What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1951–1958).

Kunz, H., & Hemelrijk, C. K. (2003). Artificial fish schools: collective effects of school size, body size, and body form. *Artificial life*, *9*(3), 237–253.

Leach, M. J., Sparks, E. P., & Robertson, N. M. (2014). Contextual anomaly detection in crowded surveillance scenes. *Pattern Recognition Letters*, *44*, 71–79.

Lee, M. (2012). *A literature review of emergency and non-emergency events*. Massachusetts, United States of America: Fire Protection Research Foundation. Retrieved from `http://books.google.com.my/books?id=T7t1kgEACAAJ`

Lekien, F., & Marsden, J. (2005). Tricubic interpolation in three dimensions. *International Journal for Numerical Methods in Engineering*, *63*(3), 455–471.

Lempitsky, V., & Zisserman, A. (2010). Learning to count objects in images. In *Advances in Neural Information Processing Systems* (pp. 1324–1332).

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

Li, T., Chang, H., Wang, M., Ni, B., Hong, R., & Yan, S. (2015). Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(3), 367–386.

Li, W. x., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(1), 18–32.

Liang, R., Zhu, Y., & Wang, H. (2014). Counting crowd flow based on feature points. *Neurocomputing*, *133*, 377–384.

Lim, M. K. (2014). *Activity understanding and abnormal event detection in video surveillance* (PhD thesis). University of Malaya.

Liu, C. (2009). *Beyond pixels: exploring new representations and applications for motion analysis* (PhD thesis). Massachusetts Institute of Technology.

Liu, C., Yuen, J., & Torralba, A. (2011). Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 978–994.

Liu, N.-H., Chiang, C.-Y., & Chu, H.-C. (2013). Recognizing the degree of human attention using eeg signals from mobile sensors. *Sensors*, *13*(8), 10273–10286.

Lloyd, C. D. (2006). *Local models for spatial analysis*. CRC Press. Retrieved from `https://books.google.com.my/books?id=bIKToJ9en1UC`

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Loy, C. C. (2010). *Activity understanding and unusual event detection in surveillance videos* (PhD thesis). Queen Mary, University of London.

Loy, C. C., Chen, K., Gong, S., & Xiang, T. (2013). Crowd counting and profiling: Methodology and evaluation. In *Modeling, simulation and visual analysis of crowds* (pp. 347–382). Springer.

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

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

MacKinnon, I. (2008). Ramadan alms-giving sparks fatal stampede in indonesia. *The Guardian*. Retrieved from `http://www.theguardian.com/world/2008/sep/15/indonesia`

Madzimbamuto, F. (2003). A hospital response to a soccer stadium stampede in Zimbabwe. *Emergency medicine journal*, *20*(6), 556–559.

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

Makris, D., & Ellis, T. (2005). Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *35*(3), 397–408.

Marana, A., Velastin, S., Costa, L., & Lotufo, R. (1997). Estimation of crowd density using image processing. In *IEE Colloquium on Image Processing for Security Applications (Digest No.: 1997/074)* (pp. 11/1–11/8).

Marana, A., Velastin, S., Costa, L., & Lotufo, R. (1998). Automatic estimation of crowd density using texture. *Safety Science*, *28*(3), 165–175.

Mazzon, R., Tahir, S. F., & Cavallaro, A. (2012). Person re-identification in crowd. *Pattern Recognition Letters*, *33*(14), 1828–1837.

McPhail, C., & McCarthy, J. (2004). Who counts and how: estimating the size of protests. *Contexts*, *3*(3), 12–18.

Mehran, R. (2011). *Analysis of behaviors in crowd videos* (PhD thesis). University of Central Florida Orlando, Florida.

Mehran, R., Moore, B. E., & Shah, M. (2010). A streakline representation of flow in crowded scenes. In *Proceeding of European Conference on Computer Vision* (pp. 439–452). Springer.

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

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors.

International Journal of Computer Vision, *60*(1), 63–86.

Moddie, M. (2004). Accidents and missed lessons. *Frontline*, *21*(16), 13–14.

Moore, B. E., Ali, S., Mehran, R., & Shah, M. (2011). Visual crowd surveillance through a hydrodynamics lens. *Communications of the ACM*, *54*(12), 64–73.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. United States of America: Harvard University Press.

Morrone, M. C., & Owens, R. A. (1987). Feature detection from local energy. *Pattern Recognition Letters*, *6*(5), 303–313.

Nedrich, M., & Davis, J. W. (2010). Learning scene entries and exits using coherent motion regions. In *Advances in Visual Computing* (pp. 120–131). Springer.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, *29*(1), 51–59.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, *69*(5), 529–541. doi: 10.1109/PROC.1981.12022

Pal, S. K., Uma Shankar, B., & Mitra, P. (2005). Granular computing, rough entropy and object extraction. *Pattern Recognition Letters*, *26*(16), 2509–2517.

Pedrycz, W. (2001). *Granular computing: an emerging paradigm* (Vol. 70). Springer Science & Business Media.

Pedrycz, W., & Bargiela, A. (2002). Granular clustering: a granular signature of data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *32*(2), 212–224.

Pedrycz, W., & Bargiela, A. (2012). An optimization of allocation of information granularity in the interpretation of data structures: toward granular fuzzy clustering. *IEEE*

Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, *42*(3), 582–590.

Popplewell, J. (1986). *Committee of inquiry into crowd safety and control at sports grounds - final report*. London, England: Her Majesty's Stationery Office. Retrieved from `http://bradfordcityfire.files.wordpress.com/2013/02/popplewell-final-report-1986.pdf`

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

Rizzi, A., & Del Vescovo, G. (2006). Automatic image classification by a granular computing approach. In *Proceedings of the 2006 16th ieee signal processing society workshop on machine learning for signal processing, 2006.* (pp. 33–38).

Rodriguez, M., Ali, S., & Kanade, T. (2009). Tracking in unstructured crowded scenes. In *IEEE International Conference on Computer Vision* (pp. 1389–1396).

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

Rodriguez, M., Sivic, J., & Laptev, I. (2012). Analysis of crowded scenes in video. *Intelligent Video Surveillance Systems*, 251–272.

Rodriguez, M., Sivic, J., Laptev, I., & Audibert, J.-Y. (2011). Data-driven crowd analysis in videos. In *IEEE International Conference on Computer Vision* (pp. 1235–1242).

Schofield, A., Mehta, P., & Stonham, T. J. (1996). A system for counting people in video images using neural networks to identify the background scene. *Pattern Recognition*, *29*(8), 1421–1428.

Seydel, R. (2009). *Practical bifurcation and stability analysis* (Vol. 5). Springer Science & Business Media.

Shah, M. (2010). Visual crowd surveillance is like hydrodynamics. In *Proceedings of the International Conference on Multimedia* (pp. 3–4).

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE*

Mobile Computing and Communications Review, *5*(1), 3–55.

Shao, J., et al. (2015). Deeply learned attributes for crowded scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4657–4666).

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Solmaz, B., Moore, B. E., & 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.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. In *Proceedings of iconference* (pp. 654–662). doi: 10.9776/14308

Still, G. K. (2000). *Crowd dynamics* (PhD thesis). University of Warwick.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. London, England: Springer-Verlag London.

Tan, B., Zhang, J., & Wang, L. (2011). Semi-supervised elastic net for pedestrian counting. *Pattern Recognition*, *44*(10), 2297–2304.

Tang, X.-Q., & Zhu, P. (2013). Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space. *IEEE Transactions on Fuzzy Systems*, *21*(5), 814–824.

Taylor, P. M. (1990). *The Hillsborough stadium disaster - final report*. London, England: Her Majesty's Stationery Office.

Thida, M., Yong, Y. L., Climent-Pérez, P., Eng, H.-l., & Remagnino, P. (2013). A literature review on video analytics of crowded scenes. In *Intelligent Multimedia Surveillance* (pp. 17–36). Springer.

Tian, Y.-l., Brown, L., Hampapur, A., Lu, M., Senior, A., & Shu, C.-f. (2008). Ibm smart surveillance system (s3): event based video surveillance system with an open and extensible framework. *Machine Vision and Applications*, *19*(5-6), 315–327.

Tilly, C. (1999). From interactions to outcomes in social movements. *How social movements matter*, 253–270.

Tribunal of Inquiry on the Fire at the Stardust, Artane, Dublin. (1981). *Report of the tribunal of inquiry on the fire at the Stardust, Artane, Dublin*. Dublin, Ireland: The Stationery Office. Retrieved from http://www.lenus.ie/hse/bitstream/10147/45478/1/7964.pdf

Tuytelaars, T. (2010). Dense interest points. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2281–2288).

United Nations, Department of Economic and Social Affairs, Population Division. (2013). *World population prospects: The 2012 revision* (Vol. 1). New York, United States of America: United Nations Publications.

United Nations, Department of Economic and Social Affairs, Population Division. (2014). *World urbanization prospects: The 2014 revision, highlights*. New York, United States of America: United Nations Publications.

Valera, M., & Velastin, S. A. (2005). Intelligent distributed surveillance systems: a review. In *IEE Proceedings-Vision, Image and Signal Processing* (Vol. 152, pp. 192–204).

Vedaldi, A., & Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia* (pp. 1469–1472).

Wang, B., Li, W., Yang, W., & Liao, Q. (2011). Illumination normalization based on weber's law with application to face recognition. *IEEE Signal Processing Letters*, *18*(8), 462–465.

Wang, X., Tieu, K., & Grimson, E. (2006). Learning semantic scene models by trajectory analysis. In *Proceeding of European Conference on Computer Vision* (pp. 110–123). Springer.

Weiss, J. (2013). *How reporters can estimate the number of people in a crowd.* Retrieved from https://ijnet.org/en/blog/how-reporters-can-estimate-number-people-crowd

World Health Organization (WHO). (2008). *Communicable disease alert and response for mass gatherings: Technical workshop*. Geneva, Switzerland: WHO.

Wu, S., & San Wong, H. (2012). Joint segmentation of collectively moving objects using a bag-of-words model and level set evolution. *Pattern Recognition*, *45*(9), 3389–3401.

Wu, S., Yu, Z., & Wong, H.-S. (2009). Crowd flow segmentation using a novel region growing scheme. In *Pacific-Rim Conference on Multimedia: Advances in Multimedia Information Processing* (pp. 898–907). Springer.

Yan, J., & Pollefeys, M. (2006). A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proceeding of European Conference on Computer Vision* (pp. 94–106). Springer.

Yao, J. T., Vasilakos, A. V., & Pedrycz, W. (2013). Granular computing: perspectives and challenges. *IEEE Transactions on Cybernetics*, *43*(6), 1977–1989.

Yao, Y. (2000). Granular computing: basic issues and possible solutions. In *Proceedings of the 5th Joint Conference on Information Sciences* (Vol. 1, pp. 186–189).

Yao, Y. (2005). Perspectives of granular computing. In *IEEE International Conference on Granular Computing* (Vol. 1, pp. 85–90).

Yao, Y. (2009). Interpreting concept learning in cognitive informatics and granular computing. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *39*(4), 855–866.

Yi, S., Li, H., & Wang, X. (2015). Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3488–3496).

Zadeh, L. A. (1996). Key roles of information granulation and fuzzy logic in human reasoning, concept formulation and computing with words. In *Proceedings of the IEEE International Conference on Fuzzy Systems* (Vol. 1, pp. 1–1).

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems* (pp. 1601–1608).

Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A., & Xu, L.-Q. (2008). Crowd analysis: a survey. *Machine Vision and Applications*, *19*(5-6), 345–357.

Zhang, Z., & Li, M. (2012). Crowd density estimation based on statistical analysis of local intra-crowd motions for public area surveillance. *Optical Engineering*, *51*(4).

Zhen, W., Mao, L., & Yuan, Z. (2008). Analysis of trample disaster and a case study–Mihong bridge fatality in china in 2004. *Safety Science*, *46*(8), 1255–1270.

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

Zhou, D., Weston, J., Gretton, A., Bousquet, O., & Schölkopf, B. (2004). Ranking on data manifolds. *Advances in Neural Information Processing Systems*, *16*, 169–176.

Zhu, X., Liu, J., Wang, J., Li, C., & Lu, H. (2014). Sparse representation for robust abnormality detection in crowded scenes. *Pattern Recognition*, *47*(5), 1791–1799.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Kok, V. J., & Chan, C. S. (2016). GrCS: Granular computing based crowd segmentation. *IEEE Transactions on Cybernetics*, 1-12. doi: 10.1109/TCYB.2016.2538765

Kok, V. J., Lim, M. K., & Chan, C. S. (2016). Crowd behavior analysis: A review where physics meets biology. *Neurocomputing*, *177*, 342 - 362. doi: 10.1016/j.neucom .2015.11.021

Lim, M. K., *Kok, V. J.*, Loy, C. C., & Chan, C. S. (2014). Crowd saliency detection via global similarity structure. In *International Conference on Pattern Recognition* (pp. 3957–3962). (Lim and Kok contributed equally)