



# University of HUDDERSFIELD

## University of Huddersfield Repository

Hao, Yu

Crowd Abnormal Behaviour Detection and Analysis

### Original Citation

Hao, Yu (2019) Crowd Abnormal Behaviour Detection and Analysis. Doctoral thesis, University of Huddersfield.

This version is available at <http://eprints.hud.ac.uk/id/eprint/35100/>

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: [E.mailbox@hud.ac.uk](mailto:E.mailbox@hud.ac.uk).

<http://eprints.hud.ac.uk/>

# **Crowd Abnormal Behaviour Detection and Analysis**

Yu Hao

Submitted for the Degree of

Doctor of Philosophy

From the University of Huddersfield

*University of*  
**HUDDERSFIELD**  
Inspiring global professionals

School of Computing and Engineering

University of Huddersfield

Queensgate, Huddersfield, HD1 3DH

March 2019

## Copyright Statement

---

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and he has given The University of Huddersfield the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

II. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

III. The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

## Acknowledgements

---

I would like to thank my supervisor, Professor Zhijie Xu. He provided me advices, inspirations and patiently instructions during this PhD program with the length of five years. I'm impressed by his persistence and responsibility of helping his students, and will take him as an example as a lecturer of Xi'an University of Posts and Telecommunications in China.

I would like to thank Professor Ying Liu and Professor Jiulun Fan in China, as well as Xi'an University of Posts and Telecommunications for providing the funding to support this program.

I would also like to thank my wife and parents for taking care of our newly born son while I'm at UK.

## Abstract

---

The analysis and understanding of abnormal behaviours in human crowds is a challenging task in pattern recognition and computer vision. First of all, the semantic definition of the term “crowd” is ambiguous. Secondly, the taxonomy of crowd behaviours is usually rudimentary and intrinsically complicated. How to identify and construct effective features for crowd behaviour classification is a prominent challenge. Thirdly, the acquisition of suitable video for crowd analysis is another critical problem.

In order to address those issues, a categorization model for abnormal behaviour types is defined according to the state-of-the-art. In the novel taxonomy of crowd behaviour, eight types of crowd behaviours are defined based on the key visual patterns. An enhanced social force-based model is proposed to achieve the visual realism in crowd simulation, hence to generate customizable videos for crowd analysis. The proposed model consists of a long-term behavior control model based on A-star path finding algorithm and a short-term interaction handling model based on the enhanced social force. The proposed simulation approach produced all the crowd behaviours in the new taxonomy for the training and testing of the detection procedure. On the aspect of feature engineering, an innovative signature is devised for assisting the segmentation of crowd in both low and high density. The signature is modelled with derived features from Grey-Level Co-occurrence Matrix. Another major breakthrough is an effective approach for efficiently extracting spatial temporal information based on the information entropy theory and Gabor background subtraction. The extraction approach is capable of obtaining the texture with most motion information, which could help the detection approach to achieve the real-time processing.

Overall, these contributions have supported the crucial components in a pipeline of abnormal crowd behaviour detecting process. This process is consisted of crowd behaviour taxonomy, crowd video generation, crowd segmentation and crowd abnormal behaviour detection. Experiments for each component show promising results, and proved the accessibility of the proposed approaches.

## List of Publications

---

- [1] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. Unsupervised Pedestrian Sample Extraction for Model Training. In the 13th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing, 2019
- [2] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. Crowd Synthesis Based on Hybrid Simulation Rules for Complex Behaviour Analysis. In Proceedings of the 24th International Conference on Automation & Computing, 2018
- [3] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. A Graphical Simulator for Modeling Complex Crowd Behaviours. In 22nd International Conference Information Visualization, 2018
- [4] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. An Innovative Crowd Segmentation Approach based on Social Force. The Twelfth International Conference on Advanced Engineering Computing and Applications in Sciences, 2018
- [5] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. Extracting Spatio-temporal Texture Signatures for Crowd Abnormality Detection. In Proceedings of the 23th International Conference on Automation & Computing, 2017
- [6] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. An Effective Video Processing Pipeline for Crowd Pattern Analysis. In Proceedings of the 23th International Conference on Automation & Computing, 2017
- [7] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. Effective Crowd Anomaly Detection through Spatio-Temporal Texture Analysis. In International Journal of Automation and Computing, 2018
- [8] Yu Hao, Zhijie Xu, Ying Liu, Jing Wang, Jiulun Fan. An Approach to Detect Crowd Panic Behaviour using Flow-based Feature. In Proceedings of the 22th International Conference on Automation & Computing, 2016

## List of Symbols & Abbreviations

---

ASM	Angular Second Moment
BOW	Bag of Word
EM	Expectation Maximization
FCNN	Fully Convolutional Neural Network
GLCM	Grey Level Co-occurrence Matrix
GAF	Group Attraction Force
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradient
HOT	Histogram of Oriented Tracklets
HS	Horn and Schunck Optical Flow
IEP	Interaction Energy Potentials
IDM	Inverse Different Moment
KNN	K Nearest Neighbor
KLT	Kanade-Lucas-Tomasi Feature Tracker
LDA	Latent Dirichlet Allocation
LBP	Linear Binary Pattern
LSTM	Long Short-Term Memory
LK	Lucas and Kanade Optical Flow
MDT	Mixture of Dynamic Textures
MBH	Motion Boundary Histogram
NMF	Non-Negative Matrix Factorization
PSO	Particle Swarm Optimization
PLE	Principle of Least Effort
ROIs	Regions-of-Interest
SIFT	Scale-invariant feature transform
SFM	Social Force Model
STT	Spatio-Temporal Texture
STV	Spatio-Temporal Volume
SVM	Support Vector Machine

# List of Figures

---

Figure 1-1. Research field hierarchy .....	3
Figure 1-2. Examples of Behaviour Analysis on Individual Behaviours and Crowd Behaviours ....	4
Figure 1-3. The Procedure of Crowd Analysis .....	5
Figure 2-1. Taxonomy of Image Features in Computer Vision .....	10
Figure 2-2. Extraction Process of HOG Feature .....	13
Figure 2-3. Hough Transform Boundary Detection .....	14
Figure 2-4. Taxonomy of Video Based Features .....	16
Figure 2-5. (a) Moving but without Optical Flow. (b) Static but with Optical Flow .....	19
Figure 2-6. The STV Modelling and STT Extraction .....	20
Figure 2-7. Acceleration, Repulsive and Obstacle Avoidance Forces in SFM .....	21
Figure 2-8. Process of Modelling Social Force .....	23
Figure 2-9. General Procedure of Conventional Approach.....	24
Figure 2-10. Procedure of DT Behaviour Recognition .....	26
Figure 2-11. Taxonomy of Deep Learning Approaches.....	29
Figure 2-12. General Procedure of Deep Learning .....	29
Figure 2-13. Rules of Boids Behavioural Model .....	33
Figure 2-14. Sample Iterations of A Cellular Automata .....	35
Figure 2-15. Structure of Block-Based Approach.....	36
Figure 2-16. A Three-Layer Structure of Hybrid Crowd Simulation.....	37
Figure 2-17. Sequential Structured Simulation Approach .....	38

Figure 3-1. Taxonomy of Crowd Behaviours .....	43
Figure 3-2. Sub-Phenomena in Bottleneck and Fountainhead .....	44
Figure 3-3. Lane Effect of Crowd .....	45
Figure 4-1. The Proposed STT Extraction Approach .....	49
Figure 4-2. Images with Corresponding Information Entropy Value .....	52
Figure 4-3. The Influence to Gabor Kernel by Changing Values of Parameters .....	54
Figure 4-4. Procedure of Boundary Detection with Gabor Filter.....	54
Figure 4-5. Results of each Processing Phase in Gabor Boundary Detection.....	55
Figure 4-6. Entropy Values of Random STTs.....	57
Figure 4-7. Selected STTs from Various Video Footages .....	57
Figure 4-8. The Parallel Lines in STT which Influence the Entropy Estimation.....	58
Figure 4-9. The Procedure of Improved STT Selection Approach .....	59
Figure 4-10. Gabor Filtering Results along Eight Directions .....	60
Figure 4-11. Comparison between Results with Eight and Six Kernel Functions .....	61
Figure 4-12. Motion Information in the STT .....	62
Figure 4-13. Crowd Panic Dispersing Detection Result .....	64
Figure 5-1. Taxonomy of Texture Extraction Approaches .....	69
Figure 5-2. GLCM Extraction.....	72
Figure 5-3. K-Nearest Neighbor Classification.....	74
Figure 5-4. Support Vectors, Decision Boundary and Interval Boundary in SVM .....	76
Figure 5-5. Structure of BP Neural Network .....	77
Figure 5-6. Structure of the Signature Modelling Approach.....	79

Figure 5-7. Trends of GLCM Patterns along Time .....	84
Figure 5-8. The Proposed Framework of Panic Crowd Behaviour Detection Approach.....	86
Figure 5-9. A Comparison of Optical Flows before and after the Neighborhood Average Procedure .....	87
Figure 5-10. Changing of Magnitude in Panic Event.....	88
Figure 5-11. Difference Values $S$ between $S$ and $A$ Value for Each Frame .....	89
Figure 5-12. Detection Results using the Proposed Framework .....	90
Figure 6-1. The Framework of Proposed Crowd Synthesis Technique .....	92
Figure 6-2. Flow Chart of A-Star Path Finding Algorithm.....	95
Figure 6-3. The Repulsive Affected by Personal Space.....	97
Figure 6-4. The Repulsive Force Affected by Relative Velocity .....	98
Figure 6-5. The Field Perception and the Impact on Result with Different Parameters .....	99
Figure 6-6. Comparison between Agents' Distribution and Extracted Grouping Centers.....	102
Figure 6-7. Comparison between Ground Truth and Segmentation Result .....	102
Figure 6-8. The Framework of the Proposed Crowd Prediction Approach .....	104
Figure 7-1. Normal and Abnormal Snapshots from UMN Dataset.....	108
Figure 7-2. Snapshots from UCSD Dataset and the Labelled Anomalies.....	109
Figure 7-3. Snapshots from Forensic Video Dataset.....	110
Figure 7-4. Snapshots from Dataset with Extreme High Density .....	111
Figure 7-5. Structure of Proposed Classification Approach.....	112
Figure 7-6. Detection Result using GLCM Signature and KNN .....	113
Figure 7-7. Detection Result using TAMURA Signature and KNN.....	114

Figure 7-8. Comparison of Detection Results on Panic Dispersing.....	115
Figure 7-9. Detection Result of the Proposed Change Detection Approach.....	117
Figure 7-10. Snapshots of Simulation Results using Proposed Approach.....	121
Figure 7-11. STTs Comparison between Simulated and Real-Life Scene.....	122
Figure 7-12. Simulated Crowd and Exhibited Motion Patterns .....	125
Figure 7-13. The Relation between Frame Rate Per-Second and Number of Agents.....	126
Figure 7-14. (a) Snapshot of the Simulated Crowd (b) The Extracted Optical Flow.....	127
Figure 7-15. (a) The Detected Edges. (b) A Comparison of Location between Ground Truth and Detected Agents .....	128
Figure 7-16. The Estimation Procedure and Prediction Result .....	130
Figure 7-17. A Performance Comparison between Proposed Pattern and Others.....	131

## List of Tables

---

Table 2-1. Pattern Comparison between Modelling Approaches.....	36
Table 3-1. Crowd Taxonomy in Various Researches .....	43
Table 3-2. Crowd Behaviours and Corresponding Patterns .....	47
Table 5-1. Comparison between Texture Patterns of Spatio-Temporal Texture Patches .....	85
Table 6-1. Segmentation Accuracy on Simulated Videos.....	103
Table 7-1. Accuracy of Multiple Signatures and Classifiers Combination.....	116
Table 7-2. STTs Pattern Value Comparison between Real-Life and Simulated Videos.....	123

# Table of Contents

---

Copyright Statement.....	I
Acknowledgements .....	II
Abstract .....	III
List of Publications.....	IV
List of Symbols & Abbreviations.....	V
List of Figures.....	VI
List of Tables .....	X
Table of Contents.....	XI
Chapter 1. Introduction.....	1
1.1. Thesis Motivation .....	1
1.2. Background.....	2
1.3. Key Challenges for Online Crowd Behaviour Analysis .....	5
1.4. Project Objectives and thesis Structure .....	6
Chapter 2. Literature Review .....	9
2.1. Image Feature Engineering.....	9
2.1.1. Color-based Features.....	11
2.1.2. Texture-based Features .....	12
2.1.3. Shape-based Features .....	14
2.1.4. Spatial-relation Features .....	15
2.2. Video Features .....	15

2.2.1 Flow-based Features .....	16
2.2.2 Spatial-Temporal Features .....	19
2.2.3 Semantic Features .....	21
2.3. Techniques for Behaviour Recognition .....	23
2.3.1. General Procedure of Conventional Approach .....	24
2.3.2. The State-of-the-art Conventional Approach.....	26
2.3.3. Behaviour Recognition using Deep Learning Framework.....	28
2.4. Crowd Simulation and Synthesis .....	32
2.4.1. Taxonomy of Crowd Simulation Approaches on Spatial Scale.....	32
2.4.2. Hybrid Crowd Simulation Models .....	36
2.5. Chapter Summary .....	39
Chapter 3. Crowd Analysis and Behaviour Modelling .....	40
3.1. Crowd Behaviour Definition and Taxonomy .....	41
3.2. Crowd Behaviour Taxonomy.....	41
3.3. Chapter Summary .....	47
Chapter 4. Spatial-Temporal Texture Feature Extraction .....	48
4.1. Baseline Operation for STT selection .....	49
4.1.1 Information Entropy .....	50
4.1.2 Boundary Detection with Gabor Filter.....	52
4.2. Implementation of Effective STT Extraction .....	56
4.2.1 Inaccurate Entropy Estimation in STT.....	56
4.2.2 Improved STT Selection Strategy using Gabor Filter .....	58

4.2.3 Computational Efficiency .....	61
4.2.4 Exploiting the STT for Panic Detection.....	62
4.3. Chapter Summary .....	65
Chapter 5. Crowd Behaviours Classification and Abnormality Detection .....	66
5.1. Image Texture Patterns.....	67
5.1.1 Taxonomy of Texture Pattern Extraction Approaches.....	68
5.1.2 Grey Level Co-occurrence Matrix .....	70
5.2. Machine Learning Classifiers.....	73
5.2.1 K-Nearest Neighbors.....	74
5.2.2 Support Vector Machine .....	74
5.2.3 Back Propagation Neural Network .....	77
5.3. Classification using GLCM.....	79
5.3.1 Modelling Features From GLCM .....	79
5.3.2 Contrast Features of GLCM.....	80
5.3.3 Orderliness Features of GLCM.....	81
5.3.4 Descriptive Statistical Features of GLCM .....	82
5.3.5 GLCM Signature Modelling .....	82
5.4. Case Study: Real-time Change Detection .....	85
5.4.1 Pre-processing and Parameter Setting.....	86
5.4.2 Feature Extraction and Post Processing .....	87
5.4.3 Signature Modeling.....	87
5.4.4 Model Training.....	88

5.4.5 Anomaly Detection .....	89
Chapter 6. Complex Crowd Behaviour Synthesis and Simulation.....	91
6.1. Hybrid Rules for Crowd Synthesis .....	91
6.1.1 Baseline Works .....	93
6.1.2 Personal Space and Relative Velocity .....	95
6.1.3 Enforced Group Social Force Model .....	98
6.2. Prediction using the Enhanced Social Force Model .....	100
6.2.1 Assumption Validation .....	101
6.2.2 Predict results on the simulated crowd.....	102
6.2.3 Structure of the behaviour prediction approach .....	103
Chapter 7. Experiments and Evaluation.....	106
7.1. Datasets for Crowd Behaviour Analysis .....	107
7.1.1 Datasets with Medium Crowd Density .....	107
7.1.2 Datasets of High Crowd Density.....	110
7.2. Classification Results using GLCM Signature.....	111
7.2.1 Training Process.....	112
7.2.2 Recognition Result.....	112
7.3. Panic Dispersing Detection Results .....	116
7.4. Game Engine based Simulation .....	118
7.4.1 Simulation Tool.....	118
7.4.2 General Installation for All Simulations .....	118
7.4.3 Installations of Different Crowd Simulations .....	119

7.4.4 Evaluation of Simulation Performance .....	121
7.4.5 Simulation of Crowd with Grouping Behaviour .....	123
7.4.6 Computational Efficiency .....	125
7.5. Crowd Prediction Result .....	126
7.5.1 Crowd Simulation for Prediction .....	127
7.5.2 Pedestrian Detection and Motion Mapping .....	128
7.5.3 Social Force Estimation .....	129
7.5.4 Evaluation of Prediction Result .....	131
Chapter 8. Conclusions and Future Work .....	132
8.1. Contributions to Knowledge .....	132
8.1.1. Explicit Crowd Behaviour Definition and Taxonomy Principles .....	132
8.1.2. Effective Spatio-Temporal Texture Extraction Approach .....	132
8.1.3. Novel Crowd Behaviour Recognition Model and Pipeline .....	133
8.1.4. Realistic Crowd Behaviour Synthesis and Prediction Methods .....	134
8.2. Future Work .....	135
Reference .....	137

# Chapter 1. Introduction

## 1.1 Thesis Motivation

In recent years, public safety has increased its importance and becoming an overwhelming issue globally. One of the major concerns is terrorism. The direct costs on human life and properties, and the potential harm to society from terrorist acts are often immeasurable. In order to tackle those challenges and to manage public safety, worldwide governments have invested enormous resources and effort into security infrastructure and techniques. For example, massive numbers of CCTV cameras have been installed in many countries across the globe. Laws and policies have been formed to authorize the usage of surveillance data for monitoring, evidence collection, and even emergency response. Furthermore, resources are invested into developing related scientific research and systems such as face recognition and tracking.

As a consequence, massive amount of video data is collected. For example, a standard CCTV camera generates around 1 GB video data per hour, assuming 10 thousand cameras are installed in a city, the daily recording of video data could add up to roughly 240 TB in size, which is impractical for long term storage and manual-based analysis. In most cases, the collected video data contains little value for post processing. In another word, these video footages contain ‘normal scenes’ only. The filtering of ‘useful’ information from these massive video footages is a difficult task. The conventional approach for obtaining required video evidence from the ocean of data is by using human operators to view all collected videos, which is extremely exhausting, inefficient, and error-prone. In order to address this issue, models and techniques for automating the filtering and detection of valuable video “events” need to be explored.

Public security research in general covers a wide range of topics, for example, public policy and regulations, scientific research and technology, and financial/social impact. This research focuses on using Computer Vision-related techniques for

improving the understanding of human crowd-based activities. The fundamental goal of this research is to effectively extract, model and recognize crowd patterns. It is envisaged that the findings and contributions of this research will be valuable for real-world problem solving and applications such as digital forensic evidence retrieval and automated abnormal behaviour alarm systems.

## 1.2 Background

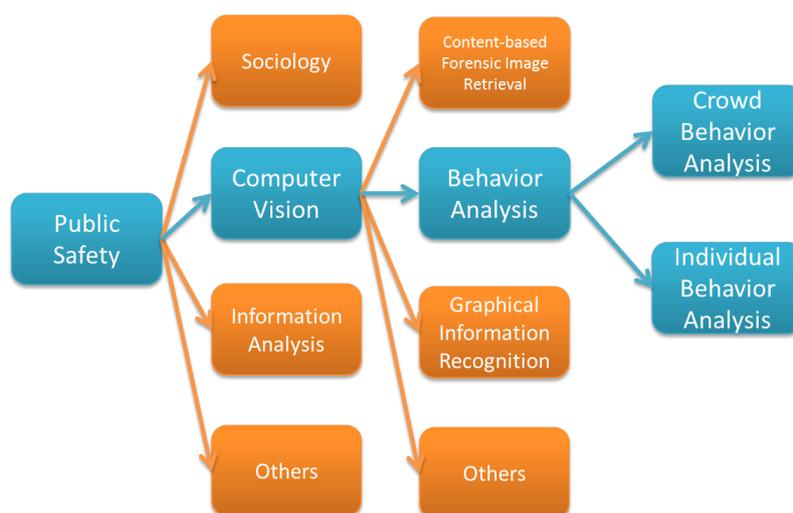
Computer Vision related researches for crowd analysis can be further categorized into application fields such as Content-based Image Retrieval (CBIR), Graphical Information Recognition (GIR) and Human Behaviour Analysis (HBA).

CBIR aims to search relative images with similar features from database according to the semantic and context of the image. It has higher requirements on retrieval speed and efficiency than conventional image retrieval that relies on pixel level operations such as histogram, moments and color set. In the last decade, some CBIR techniques attempt to utilize the semantic information extracted from an image including geometry and structure, 3D segmentation, and object recognition for various applications. Due to the nature of scientific rigorous, forensic image retrieval is a crucial tool for modern policing. Lee *et al.* (2012) has investigated the Tattoo Image Retrieval techniques that can be used by investigators as a viable way for suspect and victim identification.

GIR is also widely implemented in safety and security applications. Successful applications include fingerprint, vehicle registration plate, and shoeprint recognition technologies and systems. For example, shoeprints are capable of providing crucial evidence on suspect identification in forensic analysis. In most cases, shoeprint is a unique pattern similar to fingerprint. Even for shoes of same model, variant size and worn details can still be detected to distinguish different identities. Rathinavel and Arumugam (2011) proposed a novel approach on shoeprint recognition through matching partial shoeprint images with the whole image in a database. By using the proposed approach, the matched shoeprints can be used as valuable evidence in crime

scene investigation.

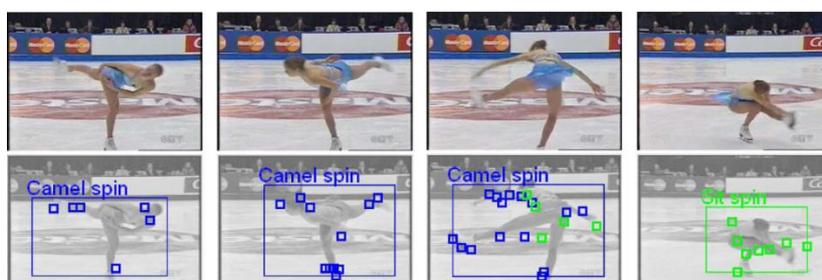
The widely installed CCTV cameras generate massive amount of live video data, these data can be utilized for predicting potential hazardous behaviours. Current main stream practical systems still reply on human operators for intervention. However, this approach has some significant disadvantages. For example, fatigue-related omission and misidentification. In order to address these issues, CV-based techniques have been developed to automate the detection tasks (Xuxin *et al.*, 2015), (Saxena *et al.*, 2008), (SangHyun and HangBong, 2014). These approaches have explicit advantages. However, the challenge is still prominent since the definition of normal and abnormal behaviours is often implicit and ambiguous in real world. Therefore, how to achieve a high detection accuracy is the problem to be addressed. Figure 1-1 illustrates the field relations of this research.



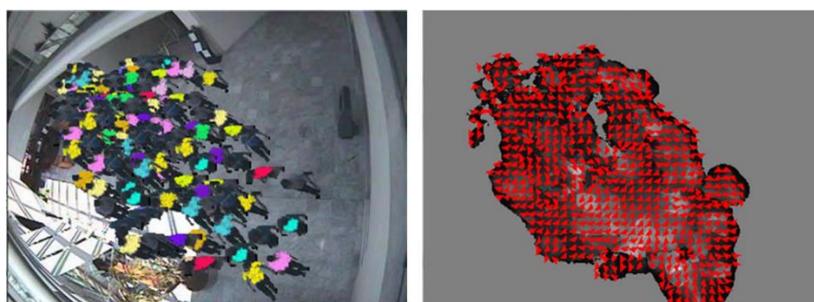
**Figure 1-1. Research field hierarchy, blocks in blue color represents the covered fields.**

HBA can be divided into individual behaviour analysis and crowd behaviour analysis. Due to the different nature of these two types, approaches are significantly different while addressing these issues. For individual behaviour analysis, the fundamental concept is to detect and verify the actual behaviour of an individual such as waving hand, walking and running. In this approach, the target individual is firstly separated from others; then defined features are extracted from the individual. The modeled feature patterns will either be used for training the behaviour templates or

classifiers, or directly used for behaviour recognition. On the contrary, for crowd behaviour analysis, individuals are often treated as a whole. For example, pedestrians of close vicinity are considered as a single entity, which is segmented from the scene; then patterns or eigenvalues are extracted from this swarm. These eigenvalues are analyzed to decide which dominant behavioural type this crowd entity belongs to. An example of individual and crowd behaviour analysis is illustrated in Figure 1-2. Figure 1-2(a) is the result of using the proposed approach in the research of Lazebnik *et al.*, (2012). In this approach, a statistical model called Bag of Word (BoW) is utilized to classify the extracted texture patch (Sivic, 2009). To be specific, the labeled image patches are firstly used to train the semantic BoW model, once the model is trained, behaviours in new images will be classified according to the texture type's statistical distribution. In Figure 1-2(a), different poses of the figure-skater such as Camel Spin and Sit Spin are detected. Figure 1-2(b) illustrates the behaviour of a crowd of people (Ernesto *et al*, 2006), where the image is segmented as foreground and background. The optical flow of the foreground is calculated to train a Hidden Markov Model (HMM) (Baum and Petrie, 1966). The HMM will then be able to decide if current scene contains abnormal crowd behaviours.



(a) Analysis of Individual Behaviours illustrated in (Lazebnik, Torralba et al)

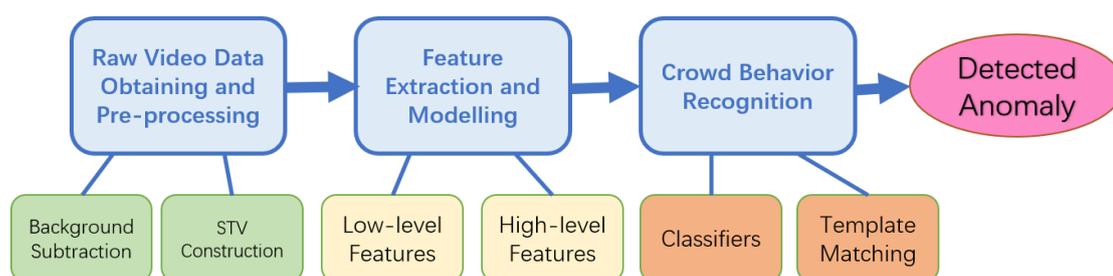


(b) Analysis of Crowd Behaviours illustrated in (Ernesto, Scott et al, 2006)

**Figure 1-2. Examples of Behaviour Analysis on Individual Behaviours and Crowd Behaviours**

### 1.3 Key Challenges for Online Crowd Behaviour Analysis

Due to the unique patterns and factors affecting human crowds, the challenges for crowd behaviour analysis are strikingly different from individual behaviour analysis. As illustrated in Figure 1-3, three main phases must be implemented to obtain the final result for online crowd analysis. In the first phase, raw video data is obtained from either video camera or simulated footage. Also, some pre-processing of the raw video data is applied such as background subtraction as the preparation of the next phase. For the second phase, crowd's low-level features are extracted from the processed video data, and further modelled into high-level semantic features. For the third phase, features are merged into descriptors to determine whether the anomaly exists in the crowd video. Various issues need to be addressed in each phase. This research managed to tackle 3 key issues in each of these three phases.



**Figure 1-3. The procedure of crowd analysis**

In the research of behaviour analysis in computer vision, wide-range of features and techniques are explored. The most appropriate feature/technique for the analysis of crowd behaviour is yet unknown. Therefore, the key issue needs to be solved is to find out the feature or descriptor which have the better performance on the recognition of crowd behaviours in the early phase of the program. The research concentrates on the support of the second and third phases of the analysis procedure to achieve the successful detection of abnormal crowd behaviours. This is the most crucial challenge to be tackled.

As the research continues, another key issue is observed while testing the devised technique for crowd abnormal behaviour classification and detection – the definition

and categorization of crowd behaviours is implicit and ambiguous. This issue greatly impacts the measurement of the proposed techniques' performance. Therefore, it needs to be addressed to support the first phase of the analysis procedure.

While training and testing the devised machine learning model for behaviour classification, another major challenge is encountered – the benchmarking video datasets for crowd behaviours are extremely limited in both quantity and quality. Therefore, the third issue is the insufficiency of the crowd video data.

## 1.4 Project Objectives and thesis Structure

In order to address the three major challenges encountered during the research, the main objectives are summarized as follows.

- Effective extraction of Spatio-Temporal Textures (STTs) for crowd behaviour representation. This objective aims to achieve the fast and effective extraction of texture with the most motion information to support the signature modelling process of crowd behaviour analysis.
- Devising a novel descriptor/technique to achieve the recognition and classification of abnormal behaviours with high performance. This objective aims to tackle the first challenge encountered, which is the ultimate goal of this research.
- Synthesis of crowd behaviours. A simulation technique has to be devised in order to produce crowd videos with desired elements. These elements include the designated crowd population, behaviours, video quality and etc.

Contributions of this research have been made as follows.

1) In the data obtaining phase, an innovative approach is devised to simulate various types of desired crowd behaviours. This contribution attempts to tackle the insufficiency of crowd videos with desired behaviours for analysis. Because the

irregular visual expression and varied form of crowd behaviours. Few comprehensive benchmarking datasets existed with all desired behavioural types. Therefore, generating crowd video using simulation algorithms become a viable and important strategy.

2) In the feature extraction phase, a STT extraction approach based on information entropy is introduced to effectively obtain texture information from the STV. The feature extraction is often the most time consumptive phase in the behaviour recognition process. This contribution attempts to obtain the STT with most motion information with the least time consumption based on the core concepts of information entropy and Gabor filtering.

3) In the behaviour recognition phase, a novel crowd behaviour descriptor based on GLCM is devised to classify abnormal behaviours such as congestion and panic dispersing of a crowd. The devised descriptor is proven to outperform classic pattern filter techniques such as TAMURA (Tamura *et al.*, 1978).

In addition, an enhanced Social Force Model (SFM) is devised to predict the individual's destination in the crowd. When pedestrians with two different destinations are mixed in a crowd, it is inherently difficult to identify the belonging of each pedestrian before the crowd is separated into two clusters. Inspired by the concept of boids model, an iterative approach is devised to estimate the pedestrian's destination through an enhanced SFM.

The structure and contents of the thesis are organized as follows:

- Chapter 2 offers a comprehensive literature review of related researches, pilot systems, and their theoretical foundations;
- Chapter 3 explores the nature of the crowd behaviours categorized by a self-defined taxonomy, which leads to the revelation of the key challenges facing modern crowd behaviour analysis approaches;
- Chapter 4 then moves on to tackle the crowd motion/behaviour representation issue by devising a texture feature extraction pipeline based on the information

entropy theory. The ultimate goal is to achieve the efficient acquirement of motion information from live video feeds;

- Chapter 5 of the thesis defines a novel descriptor based on the Grey Level Co-occurrence Matrix (GLCM) for aiding the abnormal behaviour recognition and classification;
- Chapter 6 introduces a crowd synthesis model that integrates both long-term and short-term behaviour generating mechanisms;
- In Chapter 7, experiments and evaluations are conducted on the proposed approaches of texture extraction, crowd behaviour recognition and behaviour synthesis;
- Chapter 8 concludes the research with anticipated further works.

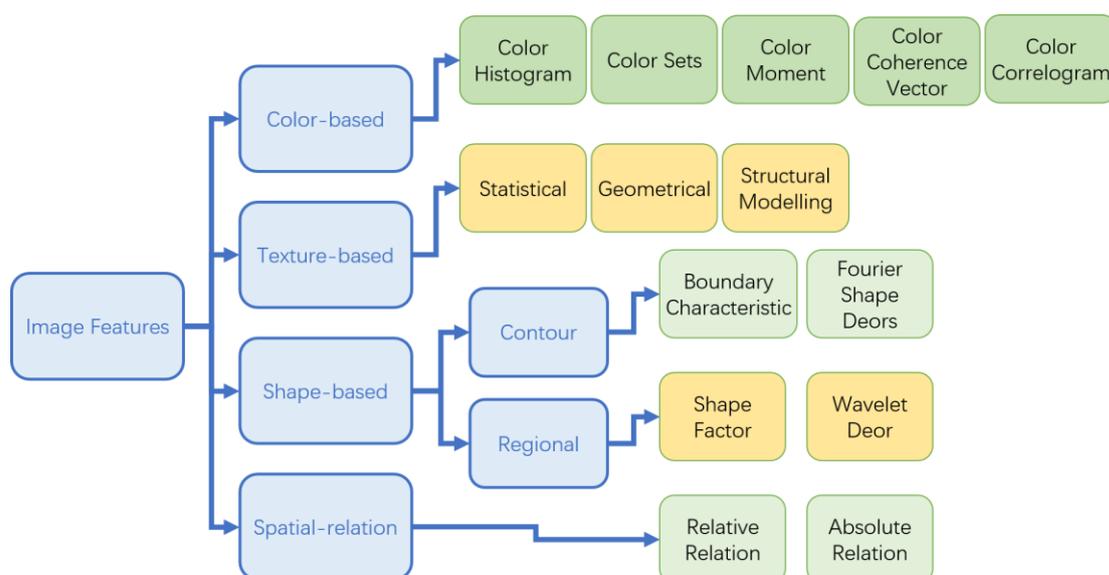
## **Chapter 2. Literature Review**

In this chapter, the relative knowledges of image and video features are reviewed. Furthermore, techniques for behaviour recognition using video data are introduced. The crowd behaviour synthesis/simulation approaches are reviewed in this chapter as well.

## 2.1 Image Feature Engineering

Video features are utilized as the primary elements to model and construct pattern detectors in this research. Since video (frame) features are usually modelled in the same manners as image features, it is useful to introduce the conventional image features first which are frequently exploited in computer vision studies.

As illustrated in Figure 2-1, conventional image features consist of four basic types, which are color-based, texture-based, shape-based and spatial-relation features respectively. Color-based feature and Texture-based feature are global-scale features, which describe the object's surface pattern in the image. Different from color-based feature, texture-based feature is often not based on single-pixel values, and could only be described by a region of neighboring pixels. Shape-based features can be further categorized into contour features and regional features. Contour features describe the boundary of a target, and regional features describe detailed information filling in an area. Spatial-relation describes the spatial or orientation relations between multiple segmented regions in an image. These relations are of adjacent, occlusion, containing etc. Spatial-relations contain relative relation and absolute relation information. The former emphasizes on the relative relations between image regions, such as up and left; while the latter focuses on the distance and angle between them.



**Figure 2-1. Taxonomy of image features in computer vision**

### 2.1.1 Colour-based features

Color-based features are global-scale features, which are often single-pixel-based. Since color-based features are insensitive to the change of target's orientation and size, they do not show good performance on capturing the local patterns. When the size of image database is too large, image indexing using color-based features will generate unwanted effects.

The Color Histogram is the most frequently used feature, it describes the global distribution of color values in the image. With advantages of rotation/translation/scaling irrelevancy, this feature can be used to describe images which are difficult to be segmented (Novak and Shafer, 1992). Since it does not include the local distribution and spatial information of colors, it is not capable of describing a specified object inside the image. Both Red-Green-Blue (RGB) and Hue-Saturation-Value (HSV) color maps are often used to quantify color histogram. The easiest approach to achieve image matching using color histogram is the Histogram Intersection. Assuming  $M(i)$  and  $N(i)$ , which are two extracted color histograms with  $k$  bins, where  $i = 1, 2, \dots, k$ . The distance of intersection  $D$  can be represented in Equation 2-1. The smaller  $D$  implies the higher similarity of two images.

$$D(M, N) = \sum_{i=1}^k \text{MIN}(M(i), N(i)) \quad 2-1$$

The Color Set is an approximation of color histogram. In order to model this feature, the color map is transformed into HSV. The transformed image is segmented and indexed with quantified color components. These color components are then transformed into a binary indexing set. Compared to color histogram, the color set feature contains spatial relation between segments. Furthermore, since the indexing set is of binary value, binary tree can be modelled to improve the indexing speed, which is beneficial in the case of large-scale image set.

In order to avoid the high dimensional vectors during the vectorization process, the feature of Color Moments is proposed in the research of Hui *et al*, (2002). An image is described using the Mean  $\mu$ , Variance  $\sigma$  and Skewness  $s$  of three vectors in color map, which can be expressed in the following Equation 2-2, 2-3 and 2-4.

$$\mu_i = \frac{1}{N} \sum_{j=1}^N P_{i,j} \quad 2-2$$

$$\sigma_i = \left( \frac{1}{N} \sum_{j=1}^N (P_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}} \quad 2-3$$

$$s_i = \left( \frac{1}{N} \sum_{j=1}^N (P_{i,j} - \mu_i)^3 \right)^{\frac{1}{3}} \quad 2-4$$

Where  $P_{i,j}$  indicates the  $i$ th color-component of  $j$ th pixel, and  $N$  is the number of pixels in the image. The moments of any components  $Y, U, V$  in color map will derive a 9-dimensional histogram vector as follow Equation 2-5. The advantage of color moment is the low dimension feature. This feature is mainly utilized for decreasing the indexing range in practice.

$$F_{color} = [\mu_Y, \sigma_Y, s_Y, \mu_U, \sigma_U, s_U, \mu_V, \sigma_V, s_V] \quad 2-5$$

The Color Coherence Vector (Greg *et al*, 1996) is another attempt to overcome the disadvantage of lacking spatial distribution of color histogram and color moment. By using a threshold, pixels in each color bin are divided into two clusters. If the size of coherence pixels is larger than threshold, these pixels are considered as converged, otherwise as non-converged. Assuming  $\alpha_i$  and  $\beta_i$  are the number of converged and non-converged pixels in  $i$ th bin, the obtained color coherence vector would be  $\langle \alpha_1 + \beta_1, \alpha_2 + \beta_2, \dots, \alpha_N + \beta_N \rangle$ .

### 2.1.2 Texture-based Features

The texture-based features are global-scale features as well. It describes the surface pattern of a target in the image. The calculation of texture-based features often involves various pixels in certain regions. Most texture-based features are rotation irrelevant, and are resistance to noise, due to their statistical nature. A common disadvantage of texture-based features is the calculating results may vary significantly when the resolution changes. Also, the changing lighting condition will misguide the successful extraction of texture features. The texture-based features can be classified as Statistical, Geometrical, and Structural based.

The Grey Level Co-occurrence Matrix (GLCM) is a typical statistical feature (Haralick *et al.*, 1973). Relationships between neighboring pixels' Grey-scale values are used to build a co-occurrence matrix, and then key features such as Energy, Homogeneity, Entropy and Correlation are modelled from GLCM. Patterns of GLCM are exploited in the following chapter of this research as well. Another statistical approach is to extract the width and orientation of texture by calculating the energy spectrum function of an image.

Histogram of Gradient (HOG) is another widely implemented statistical feature (Dalal and Triggs, 2005). It is modelled by calculating the gradient orientation's histogram of local image. HOG feature is frequently used in pedestrian detection with Support Vector Machine (SVM) (Cortes and Vapnik, 1995), since its high performance on describing the shape of object. Its extraction process is illustrated in Figure 2-2. In the process, image is transformed into Grey color map and divided into cells. For each cell the oriented gradients histogram with  $k$  bins are calculated and derived into a  $k$  dimensional vector. By using a sliding window with  $n$  cells, the HOG of this window will be normalized as a  $kn$ -dimensional descriptor. Similar statistical features include LBP (Ojala *et al.*, 2002), Haar (Viola and Jones, 2001) and SIFT (Lowe, 1999).

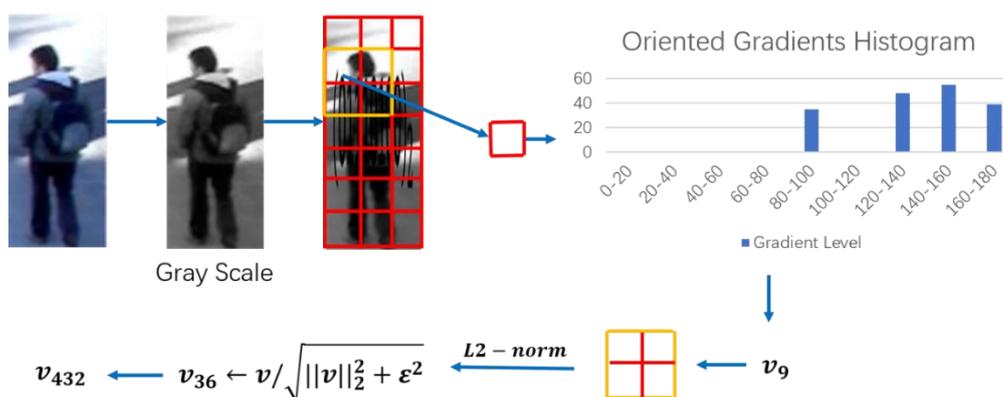


Figure 2-2. Extraction process of HOG feature

The geometrical feature is based on the assumption that complicated textures could be composed with fundamental texture elements in a certain pattern. A typical geometrical feature is the Voronoi Checkboard Pathology (Ghosh and Mallett, 1994). Another approach is using parameters of image's structural model as the texture feature.

The conventional approach of feature modelling is Conditional Random Fields (CRF) (He *et al.*, 2004), Markov Random Field (MRF) (Sean, 1996), and Gibbs Random Field (Koralov and Sinai, 2012).

In the research of Mikel *et al.* (2011), a so-called HOG3D descriptor is devised, which is modelled with the HOG feature in spatial-temporal scale. The HOG3D is used for the local behavioural matching to address the problem of detecting rare behaviours in a large dataset.

### 2.1.3 Shape-based features

Shape-based features can be divided into contour and regional features. Contour features are often extracted using boundary characteristic detection algorithms. The Hough Transform is a conventional line boundary detection approach (Richard and Peter, 1972). The process of Hough transform is illustrated in following figure. In the process, edge points in the image are firstly extracted with the Canny edge detector. Next, a table of  $\theta$  and  $r$  is modelled for each point. By transforming to the polar coordinates, the intersection of curves indicates the  $\theta_x$  and  $r_x$  of possible boundary line.

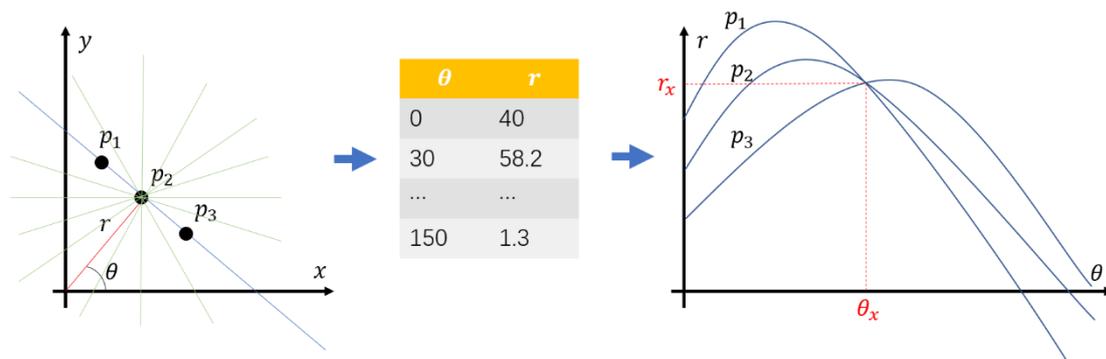


Figure 2-3. Hough Transform Boundary Detection

The Fourier Shape Descriptor is another type of contour feature (Wilhelm and Mark, 2013). Exploiting the closure and periodicity, the Fourier transform of the boundary describes the shape of object. The curvature function and centroid distance are derived from the transform as descriptors.

For regional features, shape factors such as area and perimeter are used for the

representation and matching of shapes. However, the extraction of shape factors is based on the image segmentation. The accuracy of segmentation greatly affects the performance of shape factors.

The image indexing or understanding based on shape-based features share some common problems. 1) Lacking of matured mathematical model. 2) Low adaptiveness on deformed objects. 3) Mismatching between shape-based features and human subjective observation.

### **2.1.4 Spatial-relation Features**

As above stated, the spatial-relation feature represents the spatial and directional relations between objects in the image, such as adjacent, occlusion and containing. The exploitation of spatial-relation features can improve the performance of content recognition. However, these features are sensitive to the rotation and scale-changing of the image. In practice, it is not sufficient to express the information using only the spatial-relation feature. Other features should be integrated.

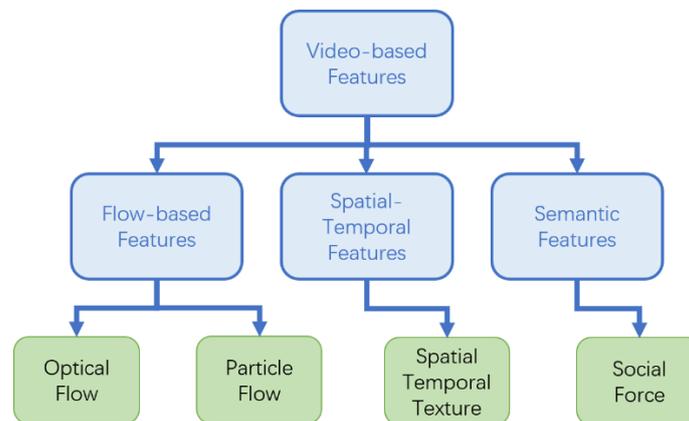
In order to extract the spatial-relation features, one approach is to segment the image into regions, then extract features and create index (Barbara and Christian, [2012](#)). Another approach is to divide the image into unified patches (Vikas *et al.*, [2011](#)).

## **2.2. Video Features**

Different from features obtained from static image, video-based features often involve temporal information. By analyzing the temporal and spatial relations among multiple or consecutive frames, features can be more accurate in revealing the nature of footages. Since each video is composed of various length of frames (somewhat equivalent to static images), conventional image features are still utilized to help the modelling of visual patterns. Unlike low-level features, video-based features usually contain high-level “event” information such as semantic expression. Therefore, video-based features exhibit higher efficiency when utilized in the research of behaviour

understanding and recognition.

As illustrated in Figure 2-4, video-based features could be classified into flow-based features and spatial-temporal features. Because the main objective of this research is to analyze the behaviour of a crowd, a social-psychological related feature Social Force (SF) is adopted into this category as well.



**Figure 2-4. Taxonomy of Video Based Features**

### 2.2.1. Flow-based features

The most frequently utilized flow-based feature is Optical Flow. Optical flow represents the instant motion of object between two consecutive video frames (Lucas and Kanade, 1981). It is often expressed as a 2-dimensional motion vector, which incorporates the velocity along  $x$  and  $y$  directions. Due to its motion information, the optical flow is widely used as a video feature in the research of behaviour recognition, motion-based segmentation, global motion compensation and video compression.

The modelling of optical flow is based on three assumptions. 1) Brightness Consistency, where magnitudes of brightness among two consecutive frames remain constant; 2) Temporal Consistency, that the time difference between two consecutive frames should be small enough to ensure the motion consistency; 3) and, Spatial Consistency, indicating that most neighboring pixels should have the identical motion.

Under ideal situation, supposing a hand gesture is captured by the camera. The spatial position of this hand is moving from one place of the first frame to another place of the next frame. At the same time, brightness magnitude of this hand remains

unchanged, which could be expressed as Equation 2-6.

$$f(x, y, t) = f(x + dx, y + dy, t + dt) \quad 2-6$$

Where  $f$  is the brightness,  $dx$  and  $dy$  are the spatial shifting,  $dt$  is the time difference between two frames. According to the Taylor expansion,  $f(x, y, t)$  could be removed and obtained Equation 2-7.

$$f_x dx + f_y dy + f_t dt = 0 \quad 2-7$$

By dividing  $dt$  on both sides we obtained Equation 2-8.

$$f_x \frac{dx}{dt} + f_y \frac{dy}{dt} + f_t = 0 \rightarrow f_x u + f_y v + f_t = 0 \quad 2-8$$

Where  $u$  and  $v$  are optical flows along  $x$  and  $y$  directions. The main stream approaches to solve  $u$  and  $v$  are the Horn-Schunk (HS) method and the Lucas-Kanade (LK) method.

#### a) Horn-Schunk method

The HS method (Horn and Schunck, 1981) adapts an additional smoothness constraint to address the solution of  $u$  and  $v$ . This constraint assumes the velocity difference between pixel and its neighbors is limited. Therefore, acquiring  $u$  and  $v$  becomes an optimization problem of the objective function.

$$\min \iint (f_x u + f_y v + f_t)^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2) dx dy \quad 2-9$$

By traversing every pixel on the image, find a combination of  $u$  and  $v$  to minimize the objective function. Its first term is the brightness constraint, and the second term is the smoothness constraint. By applying partial derivate, two functions can be obtained, which could be used to get  $u$  and  $v$ .

$$\begin{cases} (f_x u + f_y v + f_t) f_x + \lambda(\Delta^2 u) = 0 \\ (f_x u + f_y v + f_t) f_y + \lambda(\Delta^2 v) = 0 \end{cases} \quad 2-10$$

#### b) Lucas-Kanade method

The LK method made an additional assumption that the local pixels have identical optical flow (Lucas and Kanade, 1981). According to the brightness constraint equation, we can obtain the following Equations 2-11.

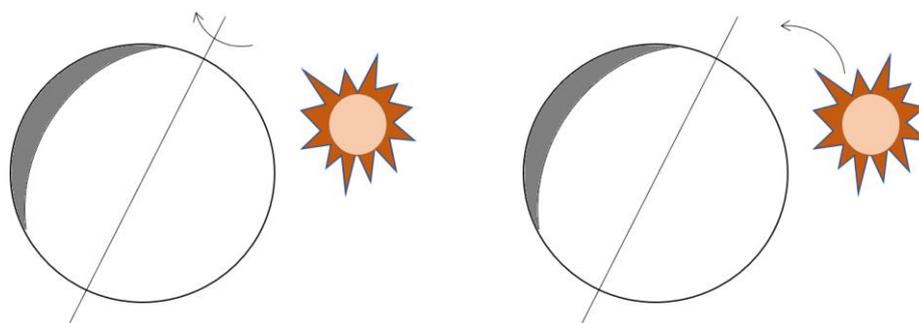
$$\begin{aligned}
 f_{x1}u + f_{y1}v &= -f_t \\
 f_{x2}u + f_{y2}v &= -f_t \\
 &\dots \\
 f_{xn}u + f_{yn}v &= -f_t
 \end{aligned}
 \tag{2-11}$$

Where  $1, 2 \dots n$  are pixels in the local window. The equations above could be represented as  $AU = b$ . By pre-multiplying the transposition of  $A$ , we obtain  $A^T A U = A^T b$ . If the inversed matrix of  $A^T A$  exists, the LK optical flow will be  $U = (A^T A)^{-1} A^T b$ .

### c) Properties of optical flow

The optical flow exhibits high efficiency on the research of behaviour recognition. After further investigation, researcher reveal following patterns of optical flow (Beauchemin and Barron, 1995): 1) The reason of its effectiveness on behaviour recognition is the invariance to the image's appearance, rather than its capability of capturing the motion information; 2) Accuracy of edges and minor spatial shifting that leads to distinctive behavioural patterns; and, 3) Object detection tasks can be achieved using optical flow information alone even on moving camera without contextual or environmental information.

The major disadvantage of optical flow is that it may not work without the constraints as stated before. For instance, it may not correctly abstract the real motion caused by objects due to illumination conditions. As illustrated in Figure 2-5, if the target object's texture is uniformed and light source is static, the rotation of spherical object will not generate interpretable optical flow. On the other hand, when the light source is moving and object remain static, the optical flow might be generated by reflections and shadows. This result implies the optical flow is sensitive to illumination conditions. When the motion velocity is too high, conventional optical flow method may be falling too. Unfortunately, objects caught in a video of high velocity are commonplace.



**Figure 2-5. (a) Moving but without optical flow. (b) Static but with optical flow.**

Various improved flow-based features are proposed by researchers to improve the robustness of optical flow, which includes Streak Flow (Mehran *et al.*, [2010](#)), Tracklet (Seung and Kuk, [2014](#)), and Particle Flow (Ali and Shah, [2007](#)).

However, in general, sparser optical flow are still confronted by aperture problems, and the denser optical flow has the disadvantage of high computational time consumption, which made it inappropriate for real-time processing.

### 2.2.2 Spatial-Temporal Features

Flow-based features can be utilized to detect dominant events, however for the unstructured high-density crowd scenes, even a fine-grain representation such as optical flow would not provide enough motion information for processing. Thus Spatial-Temporal features are used to detect abnormality to compensate the deficiencies. The related methods generally consider the motion as a whole, and characterize its spatial-temporal distributions based on local 2D pixel patches and 3D voxel (volumetric-pixel) cubic. Spatial-temporal features have sound performance in motion understanding due to their strong descriptive power (Jing and Zhijie, [2016](#)), and unlike flow-based features, the temporal information is preserved.

#### a) Spatial-Temporal Volume

The Spatial-Temporal Volume (STV) is a Spatial-Temporal model introduced by Adelson and Bergen ([1985](#)). The process of modelling the STV is illustrated in Figure 2-6. In the first step, a set of consecutive video frames is obtained from the dataset or real-time video stream. These frames will be stacked up along time sequence. The selection of the frame's indices is based on the actual video length, or the requirement

for the system. Each frame could be of Grey scale or RGB format. The size of selected data can be the entire frame or a portion of it. Once these frames are selected, they will be stacked together to generate a cube-shaped structure. According to the selection of width, height and number of frames, the size of this cube may be varied.

The STV block can be viewed as a stretching of 2D pixels into 3D voxels and filling the entire 3D space. Therefore, the STV holds certain advantages comparing to the two-dimensional patterns on behaviour analysis. For example, the STV block contains temporal information such as trajectories of pedestrians.



**Figure 2-6. The STV modelling and STT extraction**

STV is widely utilized in action recognition, for example, Bolles *et al.* (1987) introduced a technique for the recovery of the geometric and structure information from a static scene using the STV. Kühne *et al.* (2001) exploits STV to achieve 3D scene segmentation. The two latest deep learning frameworks - the 3D convolution network with STV data (Bo *et al.*, 2012) and the two-stream network with optical flows – have also been based on STV features.

### **b) Spatial-Temporal Texture**

As stated in previous section, flow-based features have sound performance when the motion velocity is low. In order to extract the temporal information in a more robust way, the detailed information can be extracted from the STV.

The textures within STV is named Spatial-Temporal Texture (STT), which could be used to extract patterns for modeling spatial and temporal signatures for behaviour

analysis. According to the requirements for different analyzing approaches, STT can be extracted in either 2D format or volume-based. The process of 2D STT extraction is illustrated in Figure 2-6. A STV block is sliced along horizontal or vertical directions. Each slice with one pixel in thickness is the obtained 2D STT. One possible approach to achieve the volume-based extraction is to implement background subtraction for each frame, and then keep the foregrounds pixels of interested objects.

In practice, researchers attempted to detect the single person's gestures such as cuts, wipes and waving hands using STT patterns (Ngo *et al.*, 1999). The approach first extract STT that is convolved with a derivative Gaussian filter. The convolved result is further processed with the Gabor decomposition where the real components of multiple spatial-frequency channel envelopes are obtained before being modelled into texture feature vectors. Finally, a Markov energy-based image segmentation approach is applied on these vectors to determine the matched gesture type.

### 2.2.3 Semantic Features

By further exploring the low-level features such as optical flow, semantic features can be modelled by incorporating “meaningful” high-level information. The most widely adapted semantic features in the research of crowd behaviour is the so-called “Social Force” for its sociology nature of describing agents behaviour in a crowd.

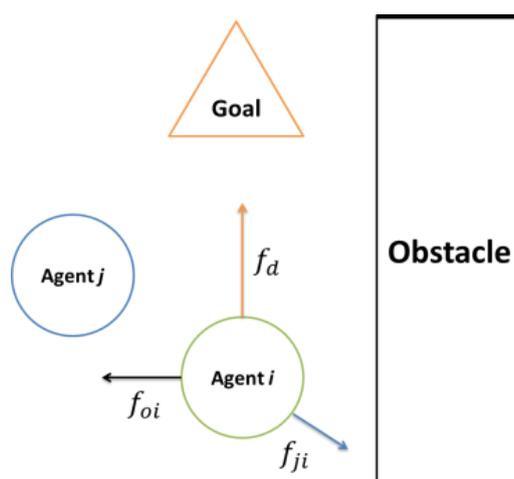


Figure 2-7. Acceleration, Repulsive and Obstacle Avoidance Forces in SFM

The Social Force Model (SFM) was first proposed by Helbing and Peter (1995). It

assumes that an agent in the crowd is always influenced by several forces derived from planned destination, neighbors in vicinity, and obstacles ahead and around. These forces determine the instant motion of a target agent. SFM can be obtained using Equation 2-12.

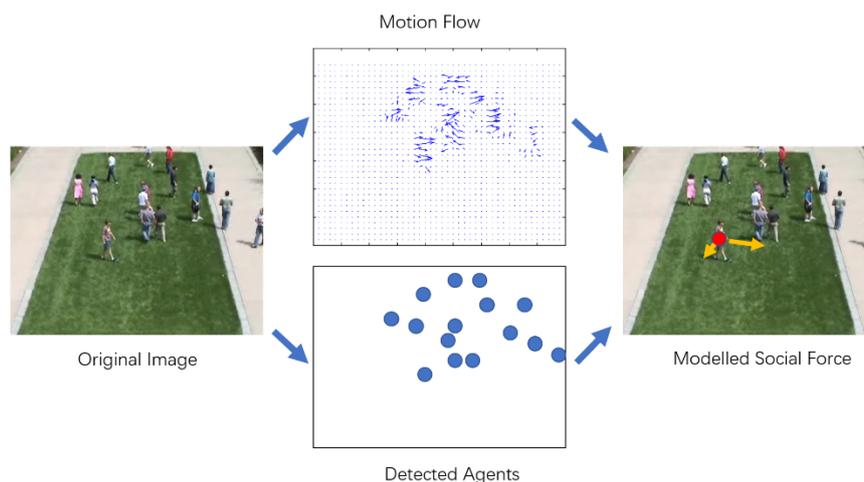
$$sf_i = f_d + \sum_{j \neq i} f_{ji} + \sum f_{o_k i} \quad 2-12$$

Where  $sf_i$  indicates the final social force applied to pedestrian  $i$ .  $f_d$  indicates the acceleration force, which derives from the agent's desire of heading the destination with constant magnitude.  $f_{ji}$  indicates the social attraction and repulsive forces, which derives to avoid collision between agent  $i$  and  $j$ .  $f_{o_k i}$  indicates the obstacle avoidance force between obstacle  $k$  and pedestrian  $i$ .

$f_{ji}$  can be expressed as Equation 2-13, where  $A_i$  and  $B_i$  are constants to control the magnitude of  $f_{ji}$ .  $r_{ij}$  is the summation of agent  $i$  and  $j$ 's radius.  $d_{ij}$  is the distance between  $i$  and  $j$ .  $n_{ij}$  is the normalized vector to control  $f_{ji}$ 's direction. The exponential function guarantees the fast magnification of  $f_{ji}$  when two agents are getting too close. On the contrary, if  $d_{ij}$  is large, magnitude of  $f_{ji}$  will have a fast reduction. Similarly, the obstacle avoidance force  $f_{o_i}$  is derived when agent moves to obstacles such as walls.

$$f_{ji} = A_i e^{(r_{ij}-d_{ij})/B_i} n_{ij} \quad 2-13$$

Process of modelling social force is illustrated in Figure 2-8. The map of motion flow/optical flow is extracted from the original image. Pedestrians' positions are located with pedestrian detector in parallel process. Then  $f_d$ ,  $f_{ji}$  and  $f_{o_k i}$  can be modelled using the introduced equations.



**Figure 2-8. Process of modelling Social Force**

The modelled social force is used as features in the training and testing processes of the machine learning model such as Support Vector Machine (SVM) to classify the pedestrian's behaviours (Lei Q. *et al.*, 2012). The research of Mehran (2009) also exploited social force with BoW statistical model to detect the abnormal frames in the video.

## 2.3 Techniques for Behaviour Recognition

The techniques for behaviour recognition can be divided into two main branches, conventional approaches and deep learning-base ones. The conventional approaches use selected video features to train the machine learning model and understand the behaviour types. On the opposite, the core concept of deep learning is to feed large number of video data to the convolution neural network, so that the network can extract most efficient features in an unsupervised manner. Generally, machine learning has better performance than conventional approaches in recent years, but some conventional approaches such as improved Dense Trajectories (iDT) still have significant efficiency on behaviour recognition. Section 2.3.1 introduces the general procedure of conventional approach. Section 2.3.2 provides the detail of the state-of-the-art conventional behaviour recognition approach, the dense trajectories. Section

2.3.3 gives introduction to the deep learning approaches.

### 2.3.1 General Procedure of Conventional approach

Behaviour Recognition is essentially a classification problem in computer vision. The general procedure of conventional approach consists of feature extraction, feature fusion and feature classification as illustrated in Figure 2-9.

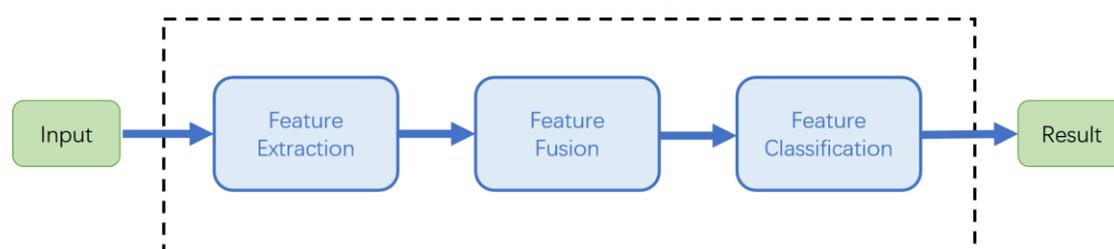


Figure 2-9. General Procedure of Conventional Approach

#### a) Feature Extraction

In order to extract features, either the 2-dimensional spatial information from image frame or the temporal information from video sequence are utilized. According to the modelling methods, the conventional feature extraction can be categorized as global and local approaches. The global approaches consider the entire video frame as an entity, extracting global feature with a top-down strategy of feature point detection, neighboring feature calculation and feature integration. Global approaches are insensitive to local occlusion and noise. On the contrary, its performance is heavily correlated to the number of feature points. The Local approaches concentrate on a segment of the video sequence, extracting feature with a bottom-up strategy of human detection, background tracking and region of interest encoding. The advantage of local approaches is the sufficiency of encoding information. The disadvantages are the reliance on the accuracy of human detection and sensitive to the noise and occlusion.

The feature points are most widely exploited feature in the conventional behaviour recognition. Some state-of-the-art features are Space-Time Interest Points (STIP) proposed by Ivan (2005), Cuboid proposed by Rabaud *et al.* (2005), Motion Energy Image (MEI) & Motion History Image (MHI) proposed by Bobick and Davis (2001) and HOG & HOF proposed by Ivan *et al.* (2008).

## b) Feature Fusion

The contour, boundary and motion features of human aren't compatible to each other. Since they usually have different dimensions and data structures, which cannot be utilized directly before being modelled. In order to acquire features with higher adaptiveness and efficiency, the feature fusion process is necessary. Furthermore, the feature fusion/encoding can remove the redundant information and enhance the accuracy of behaviour understanding. Main stream approaches of feature fusion include the Bag of Feature (BoF) and Fisher Vector. The extracted features such as texture, contour and optical flow are defined as low-level features. The fused low-level features are defined as mid-level features, and the classified features are defined as high-level semantic features in this research.

- Bag of Feature

Bag of Feature (BoF) is also named Bag of visual Word (BoW) first proposed by Lazebnik *et al.* (2006), which is originated from the Bag of Words in the linguistics. Similar to linguistics, the key features could be extracted from the image data to model the visual word. By using the k-means classifier to cluster the features, similar features are considered as one class. The center of the cluster is the visual word. The statistic of each visual word is the codebook whose size is the number of classes. Due to the differences between the text and visual features, the strategy of local feature sampling, size of codebook, weight calculation of visual word and modelling of codebook are still major challenges to be tackled.

The construction of BoW follows the order of extracting low-level features extraction and clustering, codebook establishment based on formulated clusters. And finally, classifier training using the Visual Words (clusters) from the codebook.

- Fisher Vector

Fisher Vector is another fusion technique proposed by Florent and Christopher (2007). Similar to BOW, it is also capable of achieve the normalization of feature matrices with different length. For example, the video features from videos with different length will generate feature matrices with different size. Before sending to the

neural network for classification, the feature matrices need to be processed into uniformed size. The BoW approach ignores the spatial relationship between the low-level features, and the computational complexity is high. Furthermore, the BoW needs to be retrained when a new class is inserted. The Fisher Vector addressed these issues by using Expectation Maximization to train the SIFT descriptor with a wholesale of weighting, mean and covariance matrices and spatial distribution outputs.

### 2.3.2 The State-of-the-art Conventional approach

iDT has the best performance of behaviour recognition among conventional approaches, which is proposed by Wang and Schmid (2013). In this section, iDT's original version Dense Trajectories (DT) is introduced as an example to explain the details (Wang *et al.*, 2013). The procedure is illustrated in Figure 2-10. According to the general procedure of DT, the feature points are firstly sampled with a filtering process. Next, the trajectories are extracted for each sample points. Then the STV around each trajectory is modelled for the calculation of HOG, Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH). Finally, the normalized and encoded features are classified for the analyzing result.

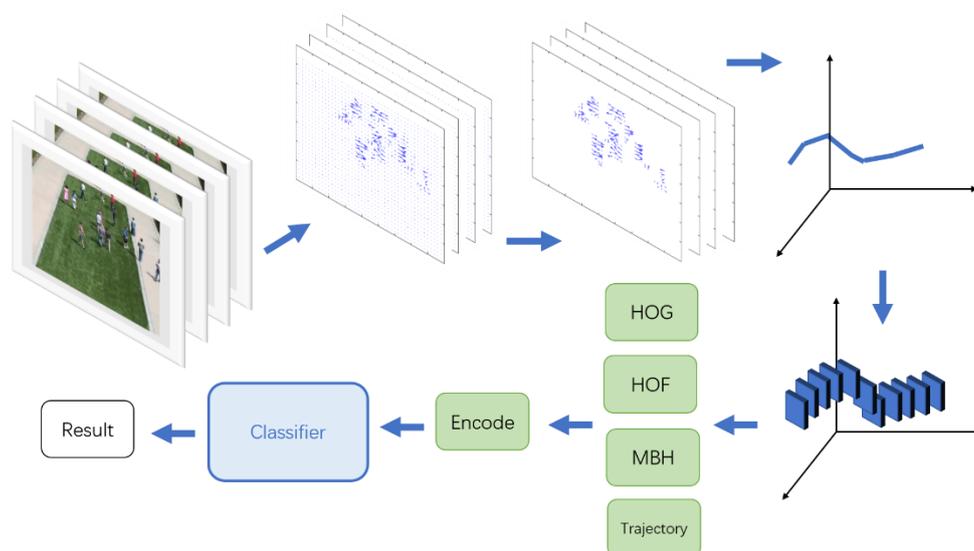


Figure 2-10. Procedure of DT behaviour recognition

#### a) Dense Feature Points Sampling

The DT extracts dense feature points in multiple spatial scales from image using

optical flow. Multiple spatial scales ensure the features to have a full coverage on all spatial scales. Next, feature points are tracked at the temporal scale. However, tracking is invalid with feature points of motionless background. Therefore, a threshold is adapted to remove feature points with low motion. The threshold  $T$  could be expressed as Equation 2-14.

$$T = 0.001 \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2) \quad 2-14$$

Where  $(\lambda_i^1, \lambda_i^2)$  are the modelled optical flow feature of pixel  $i$  in image  $I$ .  $\lambda_i^1$  indicates the flow magnitude along the horizontal direction, and  $\lambda_i^2$  indicates the magnitude along the vertical direction. The experiment indicates 0.001 is the most appropriate value in the filtering process.

### b) Trajectory Descriptor

Assume the spatial position of the extracted feature point is  $P_t = (x_t, y_t)$ , its spatial position in the next frame could be expressed as Equation 2-15.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)|_{x_t, y_t} \quad 2-15$$

Where  $w_t = (u_t, v_t)$  is the dense optical flow obtained from  $I_t$  and  $I_{t+1}$ .  $M$  is an average filter. The trajectory of a feature point during consecutive  $L$  frames could be expressed as  $(P_t, P_{t+1}, \dots, P_{t+L})$ . Because of the shifting of feature point, long-term tracking is often unreliable. Therefore, sampling process is conducted for each  $L$  frames repeatedly.

Furthermore, obtained trajectory could be used to model the trajectory shape descriptor. Assume a trajectory with length  $L$  is expressed as  $(\Delta P_t, \dots, \Delta P_{t+L-1})$ , where  $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$ . The descriptor could be obtained by regularizing with Equation 2-16.

$$D = \frac{(\Delta P_t, \dots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \quad 2-16$$

### c) Motion and Structure Descriptors

The DT and iDT further exploit features including HOF, HOG and MBH to represent the motion. Along the obtained trajectory with length  $L$ , areas with size of

$N \times N$  pixels are modelled as a STV. Next, the STV is divided by a  $n_\sigma \times n_\sigma \times n_\tau$  grid, where  $n_\sigma$  is the grid's size along spatial space, and  $n_\tau$  is the grid's size along temporal space. In DT and iDT,  $N = 32$ ,  $n_\sigma = 2$ ,  $n_\tau = 3$ . Next, HOG, HOF and MBH are extracted from these volume blocks.

- HOG – HOG is the histogram of gradient of the Grey scale image. The bin number is set as 8, thus the dimension of HOG is  $2 * 2 * 3 * 8 = 96$ .
- HOF – HOF is the histogram of optical flow including the magnitude and direction information. The number of bins is set as  $8 + 1$ , where 8 bins are identical to HOG, the extra bin is used for the statistic value of the number of pixels which is than a threshold. The dimension of HOF is  $2 * 2 * 3 * 9 = 108$ .
- MBH – MBH is the HOG feature on the optical flow map. The MBH is along two directions  $x$  and  $y$ , thus its dimension is  $2 * 96 = 192$ .

These features are normalized with the L2-norm.

#### d) Feature Encoding and Classification

For each video sequence, various trajectories could be extracted with a set of features. DT uses BOF (Feifei and Perona, [2005](#)) to encode these features. The code book is trained with 100000 feature sets, and its size is set to 4000. Once encoded, features are classified by SVM with the RBF kernel, the classification result indicates the recognized behaviour.

The framework of iDT is identical to DT, the improvement concentrates on the optimization of optical flow and feature encoding. As the result, the efficiency is significantly improved. The accuracy increases from 84.5% to 91.2% on UCF50 set (Kishore and Mubarak, [2012](#)), and from 46.6% to 57.2% on HMDB51 set (Kuehne *et al.* [2011](#)).

### 2.3.3 Behaviour Recognition using Deep Learning Framework

As the fast development of deep learning, Convolutional Neural Network (CNN) becomes the main stream classification approach in computer vision. The miss-rate of ResNet-152 is 3.5%, which outperforms the 5.1% miss-rate of human vision. In order

to achieve the fusion of spatial and temporal features, three main branches are proposed by researchers, which are Two Stream (Karen and Andrew, [2014](#)), Convolution 3D (Du *et al.*, [2015](#)) and CNN-Long Short-Term Memory (CNN-LSTM) (Tara *et al.*, [2015](#)). The structure of deep learning approaches on computer vision could be illustrated as Figure 2-11.

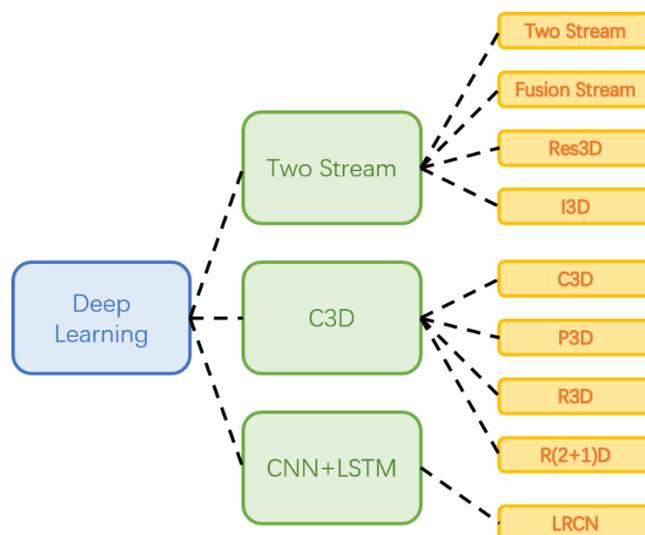


Figure 2-11. Taxonomy of Deep Learning Approaches

The general procedure of Deep Learning consists of 6 phases, which are data pre-processing, network building, definition of classification function and Loss function, definition of optimizer, training, validating and testing process. The procedure is illustrated as Figure 2-12.



Figure 2-12. General procedure of deep learning

### a) Data Pre-processing

In this phase, the data set is evenly divided as training set, validating set and testing set with random distribution. Since the deep learning is a resource consuming task, the parallel computing is required. The parallel computing includes data parallelism and model parallelism. The data parallelism is to process a mini batch of data on multiple devices, in order to achieve the parallelism of gradient computing. The model parallelism is to process different models of a neural network on different devices, in

order to reduce the training time of each iteration. The Caffe (Yangqing *et al.*, [2014](#)), TensorFlow (Martin *et al.*, [2016](#)) and Pytorch (Ketkar, [2017](#)) support data parallelism, and TensorFlow supports model parallelism.

The ordinary pre-processing of the video data includes centralization, normalization and whitening. The expansion process of the data set is required as well. The expansion is achieved by flipping, color jittering and random cropping /scaling /shifting of the image.

### **b) Network Building**

The network building process includes network architecting, network parameter initialization and overfitting handling.

For the behaviour recognition from video data, the main issue is to incorporate the temporal dimension into the network. Therefore, the conventional image recognition network architecture is adopted by networks such as VGG (Karen and Andrew, [2014](#)), GoogLeNet and ResNet (Christian *et al.*, [2015](#)). For the parameter initialization and pre-training, the Xavier or Kaming initialization is adopted. The regularization and dropout process are adopted to enhance the generalization of the model.

### **c) Classification Function and Loss definition**

The classification and regression problems are essentially identical as the primary issues of supervised learning whose learning feature is manually devised, besides the result of the former one is discrete classes and the later one is a continuous fitting curve. Therefore, the classification function could be defined with a regression function. Ordinary regression functions include Linear Regression, Logistic Regression and SoftMax Regression (Yang *et al.*, [2018](#)). Since the behaviour recognition belongs to the multi-label classification, the SoftMax Regression is most widely adopted.

Loss describes the deviation between classification result and ground truth. The goal of training process is to achieve the global or local optimal by minimizing the Loss. Regular Loss functions include Minimum Mean Square Error (MMSE) (Ephraim and Malah, [1984](#)), Hinge Loss (Junqi *et al.*, [2014](#)) and Cross Entropy Loss (Kai *et al.*, [2018](#)).

The Cross-Entropy Loss function is often exploited in the SoftMax regression.

The main stream optimizers are Stochastic Gradient Descent (SGD), SGD+Momentum, SGD+Nesterov Momentum (Bottou, [2010](#)), Adagrad, Adadelta, RMSProp and Adam (Diederik and Jimmy, [2015](#)). The Adam is usually adopted as the optimizer by default.

#### **d) Training, Validation and Testing**

In the phase of training, the data is feed to the model iteratively in the unit of mini-batch. For each iteration, the learning parameters are updated, the accuracy and loss values are returned. In the phase of validation, the performance of trained model is validated with the validation dataset for several iterations. The number of iterations is divided with K-fold approach. In the phase of testing, the testing data set is feed to the trained model to verify its generalization capability. The final accuracy could be represented with the confusion matrix, Top-1 and Top-5.

#### **e) Applications of Behaviour Recognition using Deep Learning**

In the research of Alexandre *et al.* ([2016](#)), an enforced CNN-LSTM model is proposed to predict the motion of pedestrian. In the so-called Social LSTM, each LSTM cell is provided with the information of pooled hidden state from neighboring cells. The state information is social hidden state tensor which captures the latent representation of the pedestrian, and is utilized to predict the distribution of the next possible position.

Kai introduced an enhanced structure of CNN namely Fully Convolution Neural Network (FCNN) (Kai and Xiaogang, [2014](#)). In the ordinary CNN, features of small image patches are extracted to train the model. The main disadvantage of ordinary CNN is that the computational complicity for real-time analysis. For the FCNN, fully connected layers are removed, and a  $1 \times 1$  kernel is placed at the last layer for label prediction on the segmentation map. The FCNN claims to have better performance on segmentation than conventional CNN.

## **2.4 Crowd Simulation and Synthesis**

The video data acquisition has been a tough challenge in the research of crowd behaviour analysis. The expected features of video footage for crowd behaviour analysis include crowd with different densities, hybrid behavioural types and varied camera perspectives. Nevertheless, it is difficult or even illegal to collect the desired crowd video in real-life. Using actors to perform the desired scenarios might be a good approach, however when the density is high, potential hazardous situations might occur.

In order to address this issue, the required crowd behaviours will be modelled using simulation algorithms in this research. Crowd simulation has the advantages of easiness to achieve the desired behaviours, high video quality and low cost. However, the most significant challenge is the visual realism of simulated video since it often suffers from the illogical movements which does not follow the common sense. In this section, taxonomy of crowd simulation techniques is introduced.

### **2.4.1 Taxonomy of Crowd Simulation Approaches on Spatial Scale**

According to the different spatial scales, crowd simulation can be generally categorized into the Macroscopic, Mesoscopic and Microscopic approaches.

#### **a) Macroscopic Simulation**

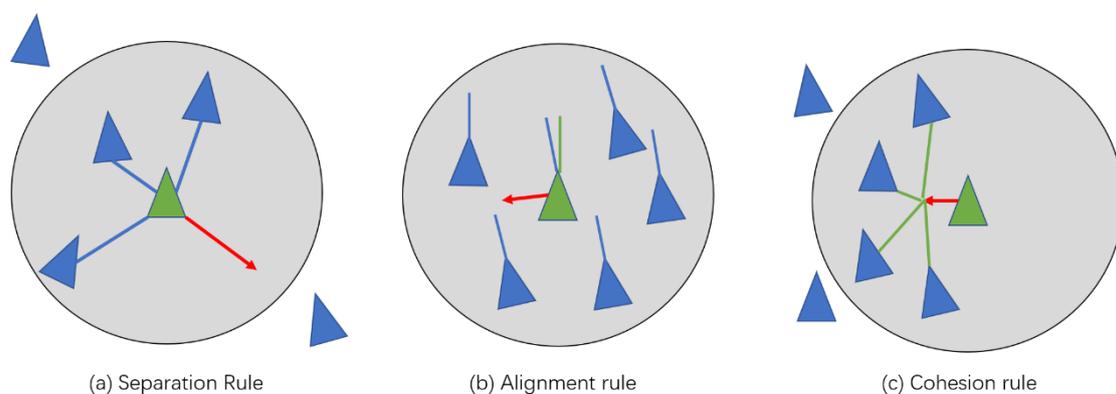
The macroscopic simulation concentrates on crowds in large size. In this approach, the modelling of the entire crowd's behaviour is the main objective rather than the behaviours of individuals. The core concept of this approach is to achieve the result by simulating the physical interaction among swarms of particles. Any social and psychological interactions between individuals are ignored. The advantage of this approach is the capability of simulating the behaviour with very large crowd scale and density. The main disadvantage is the insufficient visual realism of interactions between individuals.

#### **b) Microscopic Simulation**

The microscopic approach is most widely adapted in the research of crowd

simulation, due to its significant capability of modelling the interactions between individuals. The most representative approaches are social force model and agent behaviour model. By specifically modelling the behaviour of each individual or local motion patterns, complex crowd behaviours could be generated. And the simulated crowd size and computational time are sacrificed compared with other approaches as the price.

The detailed modelling procedure of Social Force is introduced in the section 2.2.3 of feature extraction. The concept of social force involves the psychological impact when agent interact with others among the crowd, which is originated from the Boids model proposed by Craig (1987) for the bird flock simulation. In this model, three fundamental rules are designed to control agents' movement illustrated in Figure 2-13. The first Separation rule is used as the avoidance mechanism, which prevents the collision with neighbors. In Figure 2-13(a), bird is influenced by three neighbors within its perception, and a repulsive force is derived to keep distance. The Alignment rule ensures the flock of birds to have an identical orientation with its neighbors. As Figure 2-13(b) shows, its moving direction is slightly pulled left under the influence of other's average direction within the perception. The Cohesion rule assures the bird always coherent to the flock. As shown in Figure 2-13(c), the bird is pulled to the average center of neighbors. Simulated birds mapped with Boids rules will converge as a flock. The repulsive force of SFM is derived from this concept.



**Figure 2-13. Rules of Boids behavioural model**

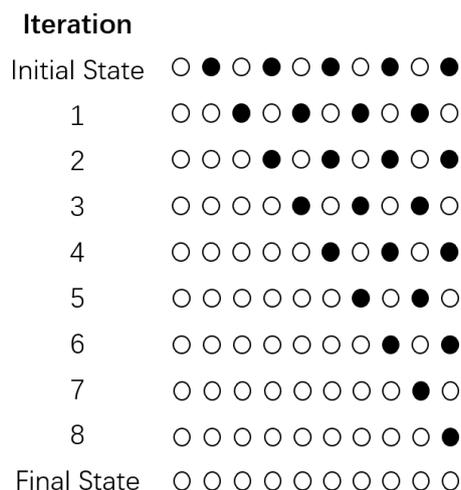
Concept of Agent-based approach is similar to SFM with some differences. In the agent-based approach, each agent is mapped with a behavioural model. Beside interaction handling among agents, this model incorporates the long-term decision making, collision handling and etc. The crowd modelled only by SFM may not handle the long-term path finding, but a well-devised agent-based model is capable of handling complicated environment. On the contrast, the agent-based approach will consume more resource than other approaches. Therefore, it isn't suitable for the simulation of crowd in high density.

### c) Mesoscopic Simulation

The mesoscopic simulation is an approach which has the performance between the macroscopic and microscopic approaches. It is capable of modeling interactions between individual at certain level, as well as maintaining the crowd's size with higher computational expense. The crowd simulation using cellular automata is a typical mesoscopic approach (Blue and Adler, [1998](#)).

The cellular automata is a discrete dynamical system, which is one of the theoretical frameworks for the complex system behavioural analysis. The definition of cellular automata is as follows. Assuming a grid of blocks is put on a  $N$ -dimensional space, each block is defined as a cell. Each cell could be in any of  $k$  states. For each iteration, every cell updates its state according to states of neighbors under the same rules.

The unique pattern of each cellular automata is determined by four main factors. 1) The motion dimension  $N$  of the cell, for example, one dimensional or three dimensional. 2) Possible states  $k$  for each cell. 3) Changing rules for each iteration. 4) The initial state of cells. Figure 2-14 illustrates the iterations of a 1-D cellular automata. Each cell has two states – black or white. The neighboring rules are as follows – for each white cell, if its left neighbor is black, then this cell will be set as black. Otherwise, this cell is set as white. The figure indicates the system becomes stable after 9 iterations.



**Figure 2-14. Sample iterations of a cellular automata**

By expanding the 1-D cellular automate into 2-D, the system can be used to simulate the crowd behaviour. Each cell represents a spatial position occupied by a pedestrian. A well devised behavioural rule determines the quality of simulation result. The mesoscopic approach improves the visual realism of individual interaction than macroscopic. However, since each cell can always be occupied by one pedestrian only, the mesoscopic approach isn't appropriate for the high crowd density.

The pattern comparison between the macroscopic, mesoscopic and microscopic approaches is listed in Table 2-1. The dilemma could be observed from the table during the selection of the appropriate approach. Using macroscopic approach for simulation, the absence of social interaction and long-term goal will hamper the visual realism. Using microscopic approach for simulation, the time consumption is not applicable for practice. Using mesoscopic approach might lose both benefits. Therefore, hybrid models are explored to devise a compromised approach which can effectively simulate the crowd in high density.

	Macroscopic	Mesoscopic	Microscopic
<b>Crowd Density</b>	High	Medium	Low
<b>Social Interaction</b>	None	None	Yes
<b>Long-term Goal</b>	None	None	Yes
<b>Time Consumption</b>	Low	Low	High

Table 2-1. Pattern Comparison between modelling approaches

## 2.4.2 Hybrid Crowd Simulation Models

The hybrid crowd simulation techniques in recent decades can be generally divided as Zone-based, Layer-based and Sequential-based. In these approaches, both macroscopic and microscopic models are adopted in different structures.

### a) Block-based Approach

For block-based approach, the footage is divided into multiple blocks. The selection of macroscopic and microscopic approaches for each block is based on the actual situation. For blocks with simple environment, the macroscopic approach is adopted to reduce resource consumption. For blocks with complicated environment which consists of diverse possible agent behaviours, the microscopic approach is adopted to provide the more accurate simulating details. The approach is illustrated as Figure 2-15.

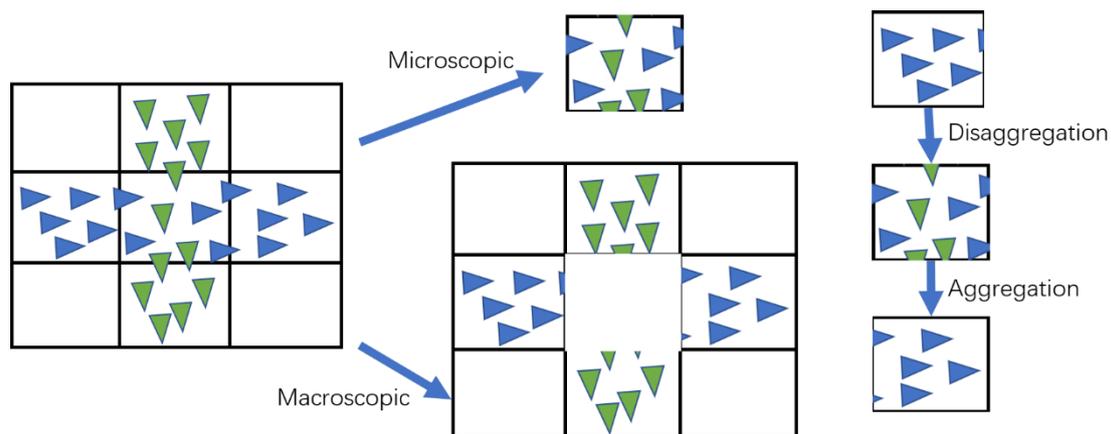


Figure 2-15. Structure of Block-based approach

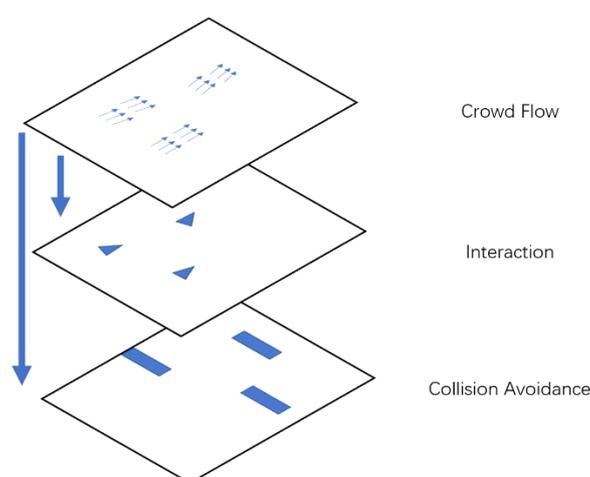
As shown in Figure 2-15, a scenario of cross road is simulated. For the intersection block which contains more agent interactions, the microscopic approach is adopted. For other blocks which has less likelihood of collision and interaction, the macroscopic is adopted to provide a general flow motion. This modeling structure is utilized in the research of Nguyen *et al.* (2011) to simulate an evacuation scenario by using the agent-based microscopic model and continuum macroscopic model. By using this hybrid approach, the simulation speed in normal blocks increases, along with the improvement of the visual realism in important blocks.

However, this kind of approach must handle the transition of agents between the macroscopic and microscopic modelled blocks. While moving from the macroscopic block to the microscopic block, agents are separated from the crowd flow using a disaggregation process. While moving on the contrast direction, an aggregation process is implemented to merge agents into a crowd flow.

Additionally, the concept of transition is expanded in the research of Sewall *et al.* (2011). In the research, the micro/macro approaches are constantly locally alternated based on features of the crowd such as the flow speed, instead of using the concept of blocks. This approach provides the switching of the simulated crowds in between global overlook and local details.

### b) Layers-based Approach

For layer-based approach, the simulation is implemented in various layers. For each layer, the general crowd flow, long-term path finding, collision avoidance and social interactions are calculated separately. The layer using macroscopic approach usually provide the general flow information of the entire crowd. Then the simulation result is feed to the microscopic layer as an input. The microscopic layer calculates the detailed activities such as social interaction with the local motion information obtained from the macroscopic layer. The structure of the layer-based approach could be illustrated as Figure 2-16.



**Figure 2-16. A three-layer structure of hybrid crowd simulation**

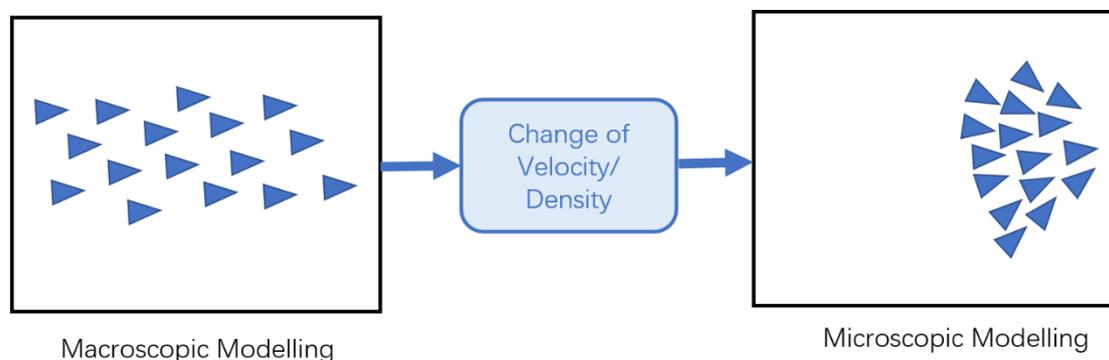
As shown in Figure 2-16, the first layer uses macroscopic approach to simulate the

general motion flow of the crowd. The interaction among agents and collision avoidance are simulated with the microscopic approaches using the flow information modelled in the first layer. In the research of Tissera *et al.* (2012), the obstacle layer is simulated with cellular automata approach. The interaction layer is simulated with social force. And the general crowd flow is simulated by scripted behavioural rules.

Comparing to the block-based, the layer-based structure doesn't avoid detailed modelling in the layer with microscopic approach. Therefore, the simulation efficiency will still be significantly impacted by the increasement of crowd size.

### c) Sequential Structured Approach

Similar to the block-based approach, the sequential structured approach also takes the concept of transferring between macroscopic and microscopic approaches. However, the transferring is temporal rather than spatial. The crowd is initially simulated in macroscopic approach to obtain the instant velocity and density. The simulating mode will be transferred under certain criteria such as the changing of local velocity. The sequential structured approach could be illustrated as Figure 2-17.



**Figure 2-17. Sequential Structured Simulation Approach.**

As shown in Figure 2-17, when the density and velocity distribution is normal, the crowd is simulated with the macroscopic modelling approach. After the change of key pattern is detected, the microscopic approach will replace the former one. In the research of Xiong *et al.* (2009), the stableness of the crowd is used as the factor to determine whether the approach should be alternated. The major shortcoming of this approach is the simulation has better performance only if the crowd is more stable. If

the simulation is unstable during most of the duration, the overall efficiency will be similar as the microscopic approach when the crowd density is high.

#### **d) Comparison between Hybrid Simulation Approaches**

From the perspective of time efficiency, the block-based hybrid approach has the best performance to others. However, this performance is based on the assumption that the situation of each block is already known. From the perspective of scalability, the approach with higher scalability always has simpler agent's behaviour simulation. The only exception is proposed in the research of Park *et al.* (2011). This sequential structured approach claims to simulate 20000 agents with interactions.

## **2.5 Chapter Summary**

In summary, this chapter reviewed the most widely utilized features and techniques for the behaviour analysis in computer vision. The image and video features are firstly introduced. Then the main stream techniques for behaviour recognition are reviewed, including both individual and crowd behaviours. Finally, the techniques for crowd simulation are introduced including the macroscopic, microscopic, mesoscopic and hybrid approaches.

## **Chapter 3. Crowd Analysis and Behaviour Modelling**

In this chapter, the primary target for this research – the Crowd – is defined. Next, a taxonomy of crowd behaviour types is induced according to the key visual features and the categorization of the state-of-the-art in related research fields. Furthermore, the key modelling techniques for each of the proposed behaviour types are introduced.

### 3.1 Crowd Behaviour Definition

By definition, a crowd is a collection of individuals with similar behaviours and physical interactions with each other. The research of social psychology indicates the general behaviour of the entire crowd does not always match the individual's behaviour. As the size of crowd increases, the individual will be influenced by the global crowd behaviour due to the loss of personal wills and control. The global crowd behaviour will be affected vice versa.

The behaviour of an individual in a crowd is determined by both physical and psychological factors. The physical factors consist of the long-term and local forces which determine the actual motion of an individual. For long-term forces, the path-finding algorithms will generate a constant driving force which “forces” the individual toward the destination. The local forces could be the social forces which decides the interaction modes. Psychological factors can also change the physical behavioural mode of an individual, such as the term “radius of comfort zone” in social force definitions. The research of Reicher and Alan (2000) introduces three-staged psychological states of an individual among the crowd including submergence, contagion and suggestion. In the submergence state, the motion of an individual is completely determined by the crowd flow while personal consciousness is temporarily removed. In the contagion state, the motion pattern of an individual is gradually affected by its neighbors. In the suggestion state, the individual acts under its own consciousness.

The global behaviour of a crowd is determined by various factors, such as the collection of individual's behaviours and the “contextual” information of an environment. Overall, the classification of crowd behaviour types is a challenging task with little consensus due to the complicated combination of various visual stimuli and models.

### 3.2 Crowd Behaviour Taxonomy

The research of Somayeh and Robert (2008) assumes that crowd behaviours are composed of basic behavioural patterns. The definition of basic behavioural patterns depends on the context and the so-called Moving Point Objects (MPO). Total 6 types of basic behaviours including Pursuit/evasion, fight, play, flock, leadership and congestion are defined. By definition, the pursuit/evasion behaviour consists of motions with high velocity and direction changing behaviours, for example, in a footage occupying large areas. The fight behaviour consists of frequently and fast physical contact between agents. The play behaviour has the hallmark of high-speed motion as well as long pauses. The flock behaviour is a crowd behaviour characterized by multiple agents with identical behaviours. The leadership behaviour is another crowd behaviour type, which consists of a leader with dominant motion patterns and a group of followers. Followers will largely remain in constant distance from the group leader. The congestion behaviour usually possesses a pattern of low velocity motion accompanied by increasing queuing time.

Hamidreza and Javad (2016) annotated the crowd activity in various datasets with both behaviour and emotion types. Five basic types of crowd behaviours are defined, including the normal state, obstacle avoidance, panic dispersing, fight and congestion. Six emotion labels are also introduced including angry, happy, excited, scared, sad and neutral. Furthermore, the authors mapped each behaviour and emotion type with specific scenarios. For example, the panic behaviours exist in scenarios such as an earthquake or terrorist attack. An emotion-based representation model of the crowd is further derived from those behaviour and emotion combinations.

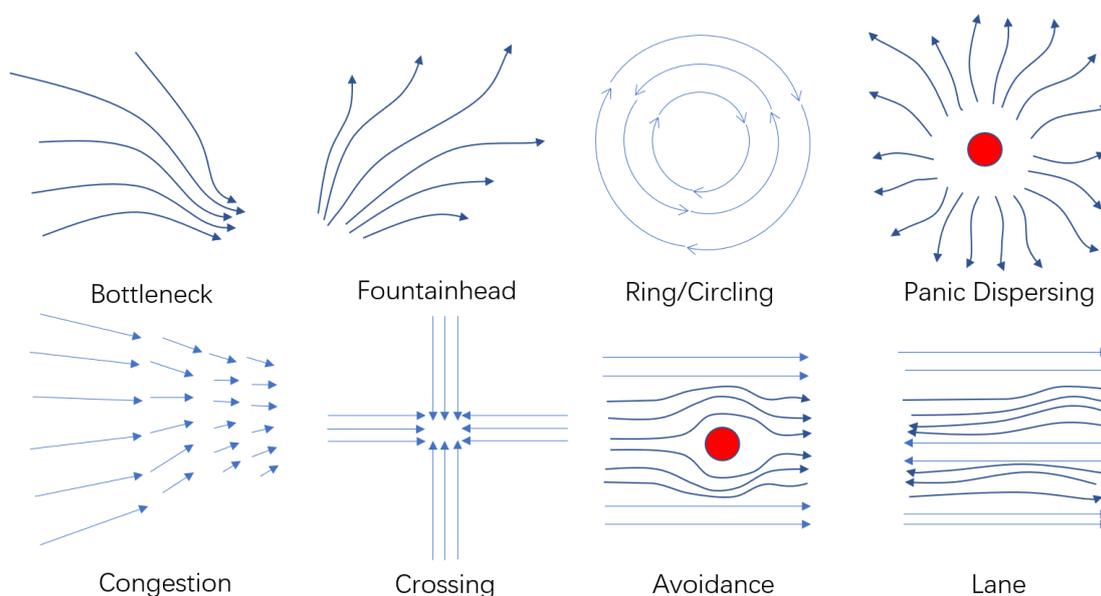
In the research of Solmaz *et al.* (2012), eigenvalues extracted from motion field are exploited to classify five different crowd behaviours. Crowd behaviours are categorized as blocking, lane, bottleneck, ring and fountainhead based on the crowd's global visual patterns.

The following Table 3-1 lists the different taxonomy of various crowd behaviours.

Research	Taxonomy
(Somayeh and Robert, <a href="#">2008</a> )	Pursuit/evasion, Fight, Play, Flock, Leadership, and Congestion
(Hamidreza and Javad, <a href="#">2016</a> )	(Physical) Normal State, Obstacle Avoidance, Panic Dispersing, Fight and Congestion (Emotional) Angry, Happy, Excited, Scared, Sad and Neutral
(Solmaz <i>et al.</i> , <a href="#">2012</a> )	Blocking, Lane, Bottleneck, Ring and Fountainhead
(Momboisse, <a href="#">1967</a> )	Casual, Conventional, Expressive, and Aggressive
(Berlonghi, <a href="#">1995</a> )	Spectator, Demonstrator, and Escaping

**Table 3-1 Crowd Taxonomy in various researches**

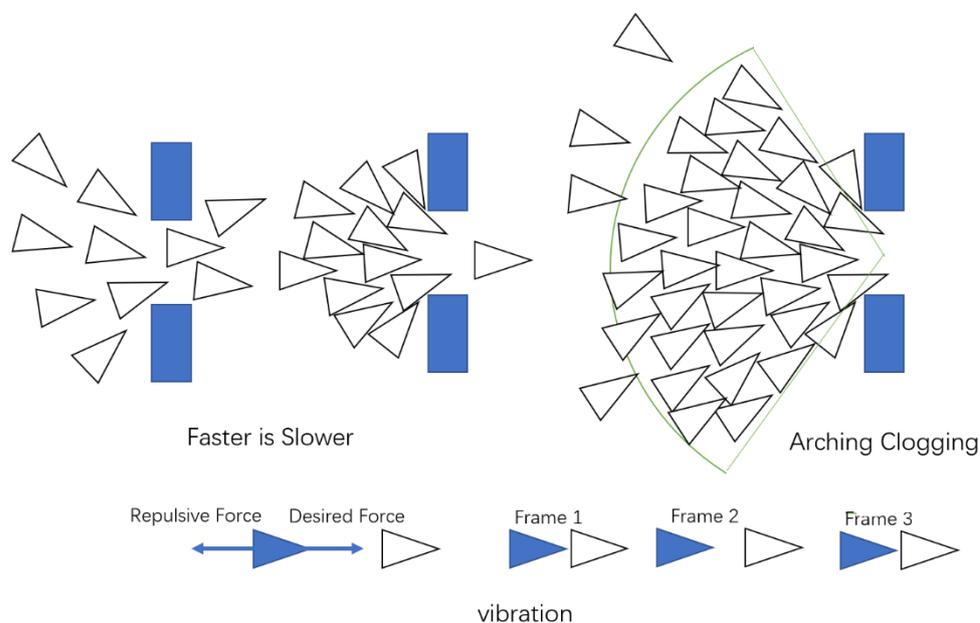
By investigating the state-of-the-art of crowd behaviour taxonomy, 8 fundamental human crowd behavioural types are proposed in this research amid their suitability for video-based analysis. According to their visual patterns, these behaviours include bottleneck, fountainhead, ring/circling, panic dispersing, congestion, crossing, avoidance, lane and the hybrid. Their global visual patterns can be illustrated as in Figure 3-1.



**Figure 3-1. Taxonomy of Crowd Behaviours**

a) Bottleneck, Fountainhead and Congestion

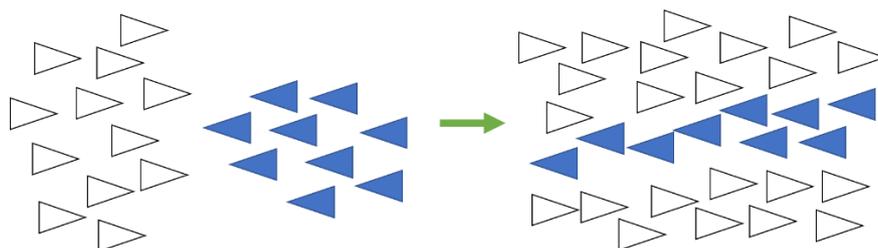
The bottleneck usually occurs when a crowd passes a narrow entrance. This crowd behaviour often exists with several intriguing sub-phenomena, which includes the faster-is-slower, arching-clogging, and oscillation effects. These effects can be exploited to evaluate the visual realism of any visual simulating attempts. For the faster-is-slower effect, individuals or agents will congest at the entrance due to the frequent collision and larger repulsive forces. On the contrast, if the crowd velocity is low, the entire gathering of agents will actually pass through the entrance in less time. For the arching-clogging effect, agents will generate an arching shape under the influence of repulsive force among agents. For the oscillation effect, if the repulsive force is not appropriately handled, vibration phenomena between agents will significantly impact the visual realism. These sub-phenomena are illustrated in Figure 3-2. On the other hand, the fountainhead crowd behaviour is a reversed circumstance of the bottleneck, which shares the similar visual patterns. The congestion is an extreme situation of bottleneck. In this behaviour, since the entrance is heavily blocked, velocities of most pedestrians are severely impacted. The farther from the entrance, the pedestrian will have higher velocity, and vice versa.



**Figure 3-2. Sub-phenomena in Bottleneck and Fountainhead**

b) Lane, Avoidance and Crossing

For bottleneck, all agents have same destination in contrast to the lane and crossing usually consists of agents moving in the opposite directions. In these circumstances, two agents have a high probability of collision. In the real-life, pedestrians will attempt to avoid collision. In order to achieve this effect in the simulation, a collision handling mechanism is mapped to the agent. The visual realism of lane effect could be utilized as a criterion of the simulation quality. In the lane/crossing scenario, agents with same direction will gradually form a lane under the influence of properly set repulsive force as illustrated in Figure 3-3.



**Figure 3-3. Lane Effect of Crowd**

The avoidance is a special case of lane formation. When a pedestrian among the crowd falls, the following pedestrians will attempt to avoid the fallen one under the influence of interaction force. Once passed by, the crowd density will become even again under the influence of interaction force as well. The crossing is also a special case of lane by generating pedestrians along four different directions. Under the influence of interaction force, the pedestrian will exhibit complex behaviours at the stage center.

#### c) Ring/Circling

For the ring and circling behaviours, agents circle around certain objects or individual with similar overall direction. The most representative of this type of behaviour is described by Hajj of Mecca (Solmaz *et al.*, 2012). When the crowd density is high, the risk of stampede increases rapidly. Also, the Ring/Circling behaviours occur at the roundabout in high density would easily trigger a traffic deadlock.

#### d) Panic Dispersing

Videos with emergent behaviours are frequently exploited to devise automatic prediction and alarming system of abnormal crowd behaviours (Xinyi *et al.*, 2011). For

crowd video containing emergent behaviours such as panic dispersing, it can normally be segmented into two subsequent stages. In the first or normal stage, emergency hasn't occurred and individuals (i.e. pedestrians) have steady motions. In the second or abnormal stage, pedestrians will attempt to run away from the disturbance source. In this stage, various particular personal behavioural patterns will be triggered and possibly combined simultaneously. The hybridization increases the difficulty in online processing. Another type of crowd behaviour is obstacle avoidance. For example, when a person falls on the ground among a moving crowd, pedestrians around will attempt to avoid the fallen person.

#### e) Hybridization

In real-life, mixtures of multiple behaviour types are widely observed. Most of researches concentrates on the singular-typed crowd abnormal behaviour detection. However, in the perspective of practical implementation, a hybrid behaviour detecting capability is at the highest demand. Due to its complexity, in this research, the hybridization of crowd behaviours is treated as an independent type. Table 3-2 summarizes the introduced crowd behaviours with their visual patterns and possible scenarios to take place.

Crowd Behaviour	Patterns	Footage
Bottleneck	Faster Is Slower Arching Clogging	Entrance
Fountainhead	Faster Is Slower Arching Clogging	Exit
Ring/Circling	Move around certain target	Roundabout
Panic Dispersing	Two Stages Escape from danger	Public area
Congested	Oscillation	Stairway
Crossing	Lane	Cross road
Obstacle Avoidance	Separate Flow	Walk path

Lane	Lane	Walk path
Hybrid	Hybrid	Complex Scenes

**Table 3-2 Crowd Behaviours and corresponding patterns**

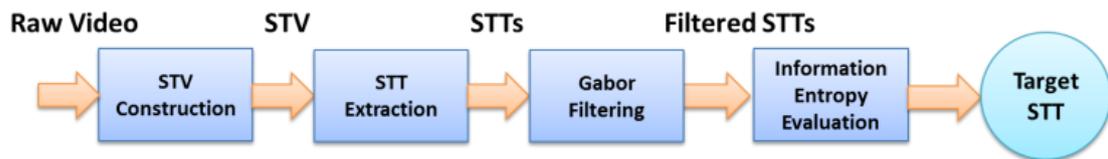
### 3.3 Chapter Summary

In this chapter, the definition of the crowd is introduced as the foundation of the research of crowd behaviour analysis. Next, taxonomies of crowd behaviour types are introduced. Based on these taxonomies, a novel taxonomy of crowd behaviour types is proposed according to their key visual features. The details of visual features are also discussed. In the experiments, crowd behaviours in the taxonomy are simulated and utilized for detection and classification.

## **Chapter 4. Spatial-Temporal Texture Feature Extraction**

In this chapter, the theoretical model and baseline approach involved in the extraction process is firstly introduced. Then a corresponding processing pipeline and its functional modules are explained. Finally, a practical test and feasibility study on the devised framework is provided.

The modelling of STV starts with stacking up consecutive frames from video without identifying any motion information within it. In order to further exploiting a 3D STV “block”, the 2D texture of, for example, a pedestrian’s motion can be extracted from it. However, the motion density within a STV is often distributed unevenly. For large portions of a STV, dynamic information can be extremely sparse or even non-exist. Therefore, one of the key problems is to obtain the texture with most abundant motion information from the STV. The main challenge is to separate the motion information from the noise such as the background. Usually each voxel (volumetric pixel - the building block of STV) will be tested to decide if it contains motion information. However, this approach is very time consuming and will greatly impact on the real-time performance of a live system. In this research, an innovated approach is proposed to effectively acquire the texture with most motion information. The procedure of this approach is illustrated in Figure 4-1.



**Figure 4-1. The proposed STT extraction approach**

The devised approach exploits the concept of information entropy to select the extracted texture slice with most abundant information. Based on the magnitude of motion trails along with the continuously evolving STV, the information entropy is obtained with a Gabor filter. The STT with the highest entropy value will be selected as target STT which will be exploited for feature extraction at the following stage. On the other hand, this approach enables the extraction of the spatial-temporal information in an efficient way in comparison to the optical flow-based calculation across the entire STV block.

## 4.1 Baseline Operation for STT selection

The proposed STT extraction approach involves the concepts of information entropy and Gabor background subtraction. The Gabor background subtraction is used to handle the unique streak-line noise in the extracted STT for a more precise entropy estimation. The information entropy lays a solid theoretical foundation for evaluating the quality of the extracted STT.

### 4.1.1 Information Entropy

The concept of information entropy is also referred as Shannon Entropy (Shannon, 1948). In the theory of thermal dynamic, the entropy increases when the information/energy losses. In the field of information theory, the data with lower probability usually contains more information. And the generation of data is a negative entropy process, therefore, the information entropy should be negative to dynamic entropy. The definition of information entropy  $H(X)$  is expressed as Equation 4-1.

$$H(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i) \quad 4-1$$

where  $x_i$  is a random variable with  $n$  possible output from probability function. The information entropy becomes larger as the uncertainty of the variable increases, more information is required to make it explicit. The information entropy has three properties as follows.

- a) Monotonicity. The event with higher probability has lower information entropy. An extreme case could be “the sun rises from the east”. Since it is a confirmed event, zero number of information is involved. From the perspective of information theory, none of the uncertainty is removed from this event.
- b) Nonnegativity. The information entropy can't be negative. Because it is impossible to increase the uncertainty by obtaining the information.
- c) Cumulativeness. The overall uncertainty of multiple random events could be expressed as the summation of each event's uncertainty. If events  $X = A$  and  $Y = B$  occurred simultaneously, where they are independent  $p(X = A, Y = B) = p(X = A) \cdot p(Y = B)$ , then information entropy  $H(A, B) = H(A) + H(B)$ .

The implementation of information entropy is further expanded to the visual information measurement. In this research, it is assumed that the image with higher entropy contains more motion information. The primary goal of this calculation is to obtain the STT with most motion (highest information entropy value).

In the phase of raw video data pre-processing, the STV is sliced with horizontal and vertical cuts along time axis. The sampling density can be a variable depending on application circumstances. The sampling density links directly with computational workload, so a balance needs to be ensured. Once the STT selection strategy is decided, the information entropy can be calculated. Overall speaking, the STTs with larger entropy value will be selected for feature extraction. To compute information entropy of each STT, it is firstly transformed from the RGB space to Grey scale. The Grey scale value is then divided into  $n$  bins. The  $x_i$  denotes the number of pixels distributed in the range of Grey scale level  $i$ .  $P(x_i)$  denotes the probability of pixels in Grey scale level  $i$  in STT.  $H(X)$  denotes the information entropy. The process is explained in the form of pseudo code in Listing 4-1.

---

```

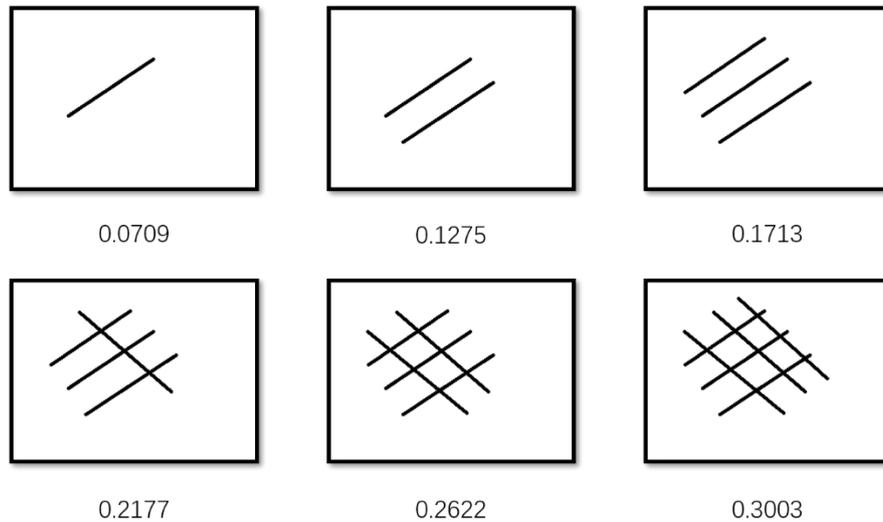
I ← STT                                % Obtain STT ↵
I ← ToGrayScale(I)                       % Transform to Gray Scale Image ↵
hist ← zeros(1,256)                       % Set histogram array with 256 bins ↵
(w, h) = size(I)                          % Obtain the width and height of image ↵
For m = 1:w                               % For each pixel in the image ↵
    For n = 1:h ↵
        i = I(m, n)                       % Obtain this pixel's gray scale value ↵
        hist(i) = hist(i) + 1             % Accumulate this bin by 1 ↵
    End ↵
End ↵
hist(i) = hist(i)./(m * n)                % Transform to probability ↵
↵
H = 0                                     % Information Entropy ↵
For k = 1:length(hist)                   % For each bin ↵
    H = H - hist(k) * log2(hist(k)) % Accumulate the entropy of each bin ↵
End ↵
return H                                 % Return Information Entropy ↵

```

---

**Listing 4-1. Pseudo Code of Calculating Information Entropy**

Figure 4-2 exhibits several sample images and their corresponding information entropy values. The assumption is proved by the calculation results in the figure: the image with more complicated patterns has higher value of information entropy. Therefore, one can further assume that the identified STT containing more motion information represented by the complex texture patterns.



**Figure 4-2. Images with corresponding Information Entropy Value**

### 4.1.2 Boundary Detection with Gabor Filter

The Gabor transform is a special short-time Fourier transform (Sejdić *et al.*, 2009). In the biological perspective, the mechanism of Gabor wavelet convolving with the image is similar to the reaction of the single vision cell to the stimulus. This mechanism is sensitive on the processing of visual data, and robust in the changing environment. Therefore, this approach has been widely exploited in the researches of object detection and background subtraction. Deepak and Meher (2015) proposed a hierarchical approach of background subtraction based on both block and pixel information using the Gabor transformed magnitude features. In order to address some disadvantages of the conventional background subtraction approach, Zhou *et al.* (2008) attempts to utilize the five-frequency circular Gabor transform to achieve a better subtraction result.

#### a) Definition of Gabor Filter

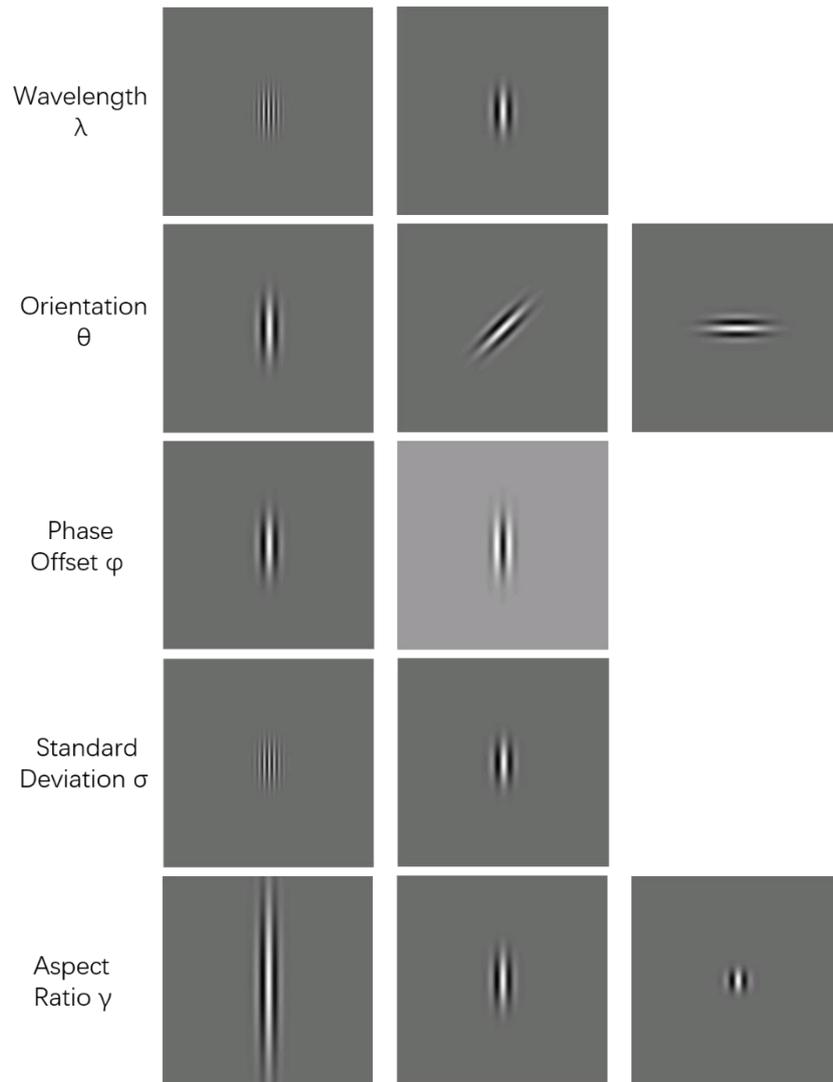
By definition, the two-dimensional Gabor kernel function is obtained by multiplying the Gaussian function and the sinusoidal function in the spatial-domain (Shen and Bai, 2004). The filter slides through entire image to implement the convolution. In practical, the Gabor window function is capable of extracting border or motion patterns along any direction. The kernel could be expressed as Equation 4-2.

$$G_{\lambda,\theta,\varphi,\sigma,\gamma}(x,y) = e^{-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}} \cos(2\pi \frac{x'}{\lambda} + \varphi) \quad 4-2$$

Where the  $x'$  denotes the size of window function along  $x$  axis, and  $y'$  denotes the size of window function along  $y$  axis. The value of  $x$  is in the range between  $-sx$  and  $sx$ . The value of  $y$  is in the range between  $-sy$  and  $sy$ .  $\lambda, \theta, \varphi, \sigma, \gamma$  are the five parameters to determine the final form of kernel. Furthermore, Equation 4-3 gives the definition of  $x'$  and  $y'$ .

$$\begin{aligned} x' &= x \cos\theta + y \sin\theta \\ y' &= y \cos\theta - x \sin\theta \end{aligned} \quad 4-3$$

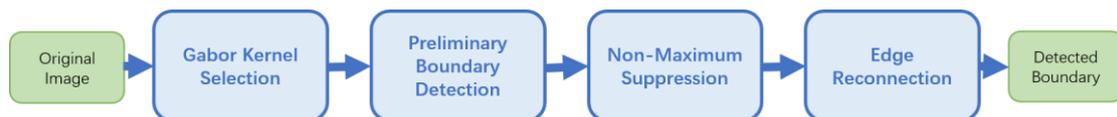
The five parameters  $\lambda, \theta, \varphi, \sigma, \gamma$  denotes the Wavelength, Orientation, Phase Offset, Standard Deviation and Aspect Ratio of the kernel function. The influence of changing these parameters is illustrated in Figure 4-3. By increasing the value of  $\lambda$  or  $\sigma$ , the cycle of kernel becomes larger. By changing the  $\theta$ , the orientation of kernel will rotate. By changing the  $\varphi$ , the kernel switches between sine and cosine functions. The  $\gamma$  controls the eclipse shape of the kernel function, large  $\gamma$  will generate a kernel with flat shape.



**Figure 4-3.** The influence to Gabor kernel by changing values of parameters

#### b) Boundary Detection

In this section, the technique of boundary detection using Gabor Filter is introduced. The procedure consists of four phases, including Gabor Kernel Selection, Preliminary Boundary Detection, Non-Maximum Suppression and Edge Reconnection as illustrated in Figure 4-4.



**Figure 4-4.** Procedure of Boundary Detection with Gabor Filter

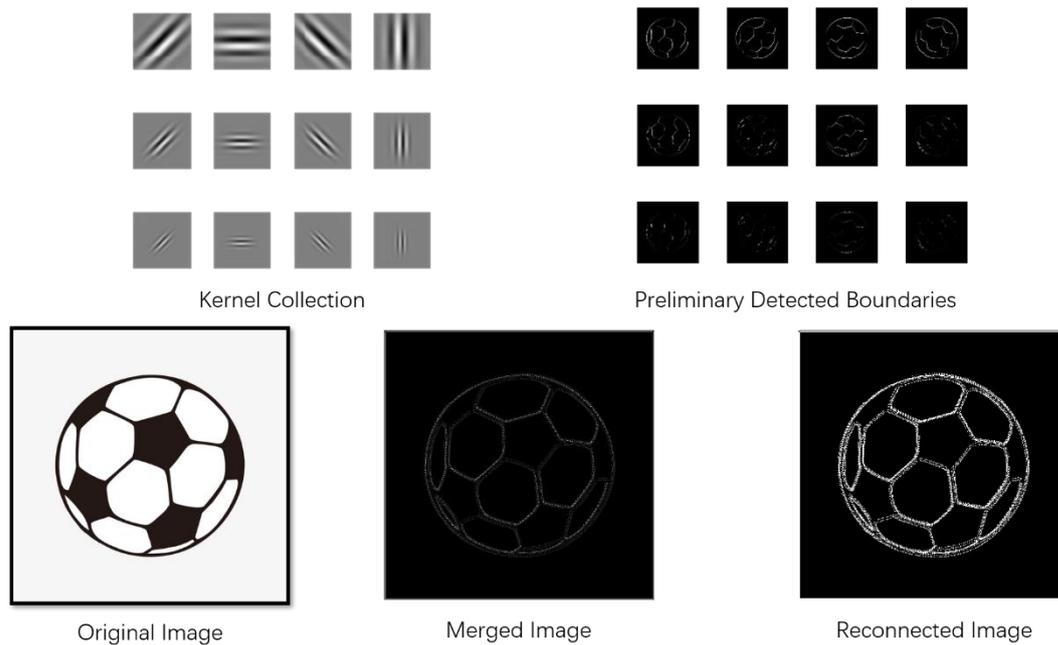
In the Gabor kernel selection phase, a collection of Gabor kernel functions is generated by setting the values of  $\lambda$  and  $\theta$ . As an empirical example, the values of  $\lambda$

are set as 6.6, 3.3 and 2.2. The values of  $\theta$  are set as 0,  $\pi/4$ ,  $\pi/2$  and  $3\pi/4$ . Therefore, a collection with 12 kernel function is obtained as illustrated in Figure 4-5.

In the preliminary boundary detection phase, the image is firstly transformed into the Grey scale colormap, then convoluted with each kernel in the collection, to obtain the preliminary boundary along each orientation.

In the non-maximum suppression phase, the Grey scale level of each pixel is compared with its neighbors along the detecting orientation. The pixel will be conserved if it is local maximum. Otherwise, it will be set as zero.

In the edge reconnection phase, the processed images will be merged. Then the remaining pixels will be reconnected into boundaries. In this example, for each pixel, neighbors in 8 orientations with Grey scale value lower than the threshold will be removed from the boundary candidates. The detected boundary is illustrated in Figure 4-5.



**Figure 4-5. Results of each processing phase in Gabor boundary detection**

### c) Extracting Motions in STT

As previously claimed, the STT with more motion information assumes to have higher information entropy. However, the static edges in the footage is likely to produce

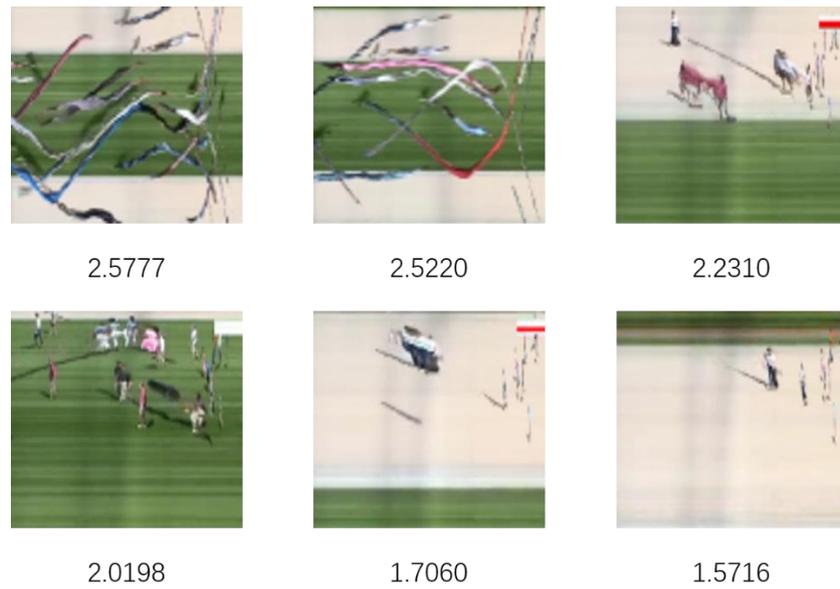
a number of entropies. Therefore, STT containing more background information may have higher entropy than the one with more motion information. In this case, the proposed STT selection mechanism based on entropy will fail. Due to the unique pattern of STT, the Gabor Boundary Detector can be utilized as a background subtractor to extract the foreground motion information. Details will be explained in the next section.

## **4.2 Implementation of Effective STT Extraction**

As introduced in Figure 4-7, a four-step procedure based on the baseline concepts and the achieved effective STT extraction technique is devised. In this section, the logic of this procedure is explained. The contents are distributed as follows. Section 4.2.1 proves the assumption that STT with more motion information has higher entropy. It also reveals the issue of inaccurate entropy estimation. Section 4.2.2 addresses the issue by importing the Gabor Boundary Detector. Section 4.2.3 further improves the Boundary Detector to achieve a better STT selection quality.

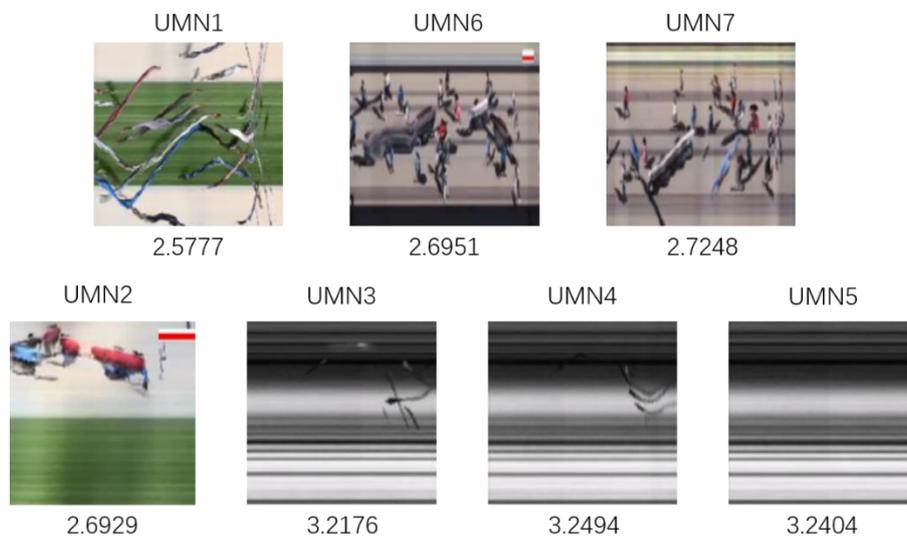
### **4.2.1 Inaccurate Entropy Estimation in STT**

It is proved in Figure 4-2 that images (STTs) with more complicated patterns often have higher entropies. Since the motion information randomly distributed in the extracted STT, the STT with more motion supposes to leave more complex trajectory patterns. As illustrated in Figure 4-6, six STTs extracted from the same STV and their corresponding entropy values are listed in descend order. The ribbon-shape patterns are the pedestrian's trajectories in STT. Visually, the STT with more ribbon-shape patterns has higher entropy.



**Figure 4-6. Entropy values of random STTs**

In order to validate this observation, the proposed STT selecting approach is applied on more video data. Seven video footages of the benchmarking dataset UMN is adopted. The detailed information of UMN is introduced in Chapter 7.1.1. 20 STTs are extracted vertical and horizontal from each modelled STV. The STT with highest entropy is illustrated in Figure 4-7. Surprisingly, only selected STTs from UMN1, UMN6 and UMN7 have the ribbon-shape motion patterns. STTs from UMN2, UMN3, UMN4 and UMN5 has high entropy, however, they contain meaningless parallel lines instead of trajectories.



**Figure 4-7. Selected STTs from Various Video Footages**

By further analyzing the incorrectly selected STT, the meaningless lines produce even higher entropy than trajectories. As illustrated in Figure 4-8, these lines are generated by stretching the static background pixels with high contrast along the time axis. The existence of these lines leads to the inaccurate entropy estimation. The intuitive thought to address this issue will be to remove these lines. As previously introduced, the Gabor filter is capable of detecting the boundary. The extraction of motion pattern's boundary is equivalent to the subtracted foreground of a STT. In next section, an approach of background subtraction using Gabor filter is introduced.

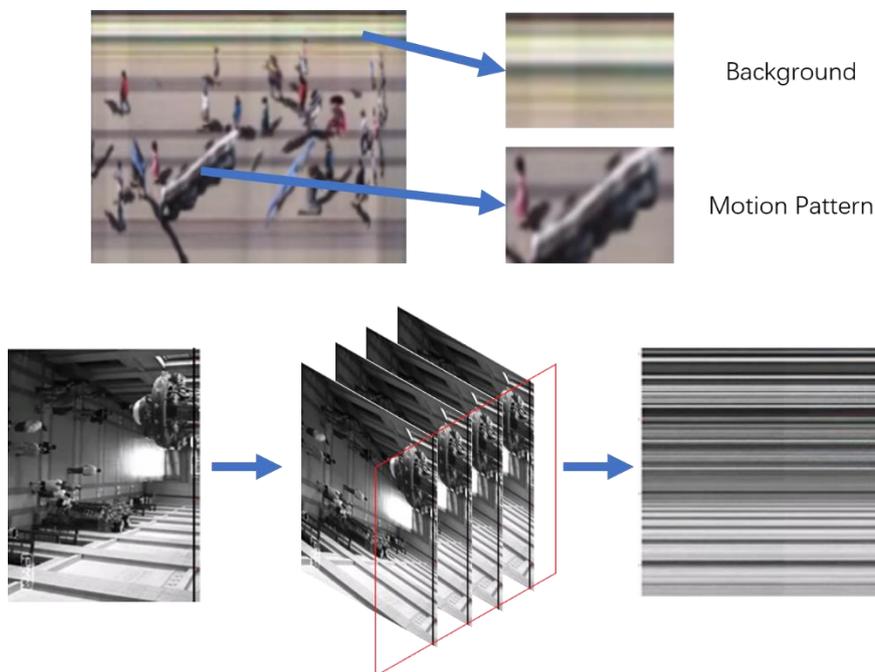
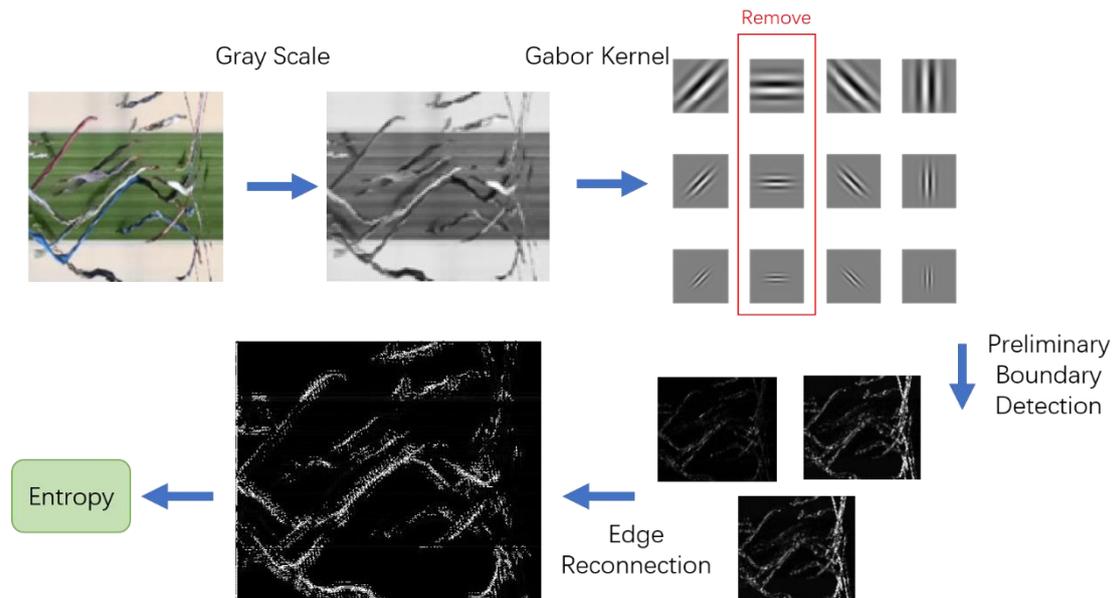


Figure 4-8. The parallel lines in STT which influence the entropy estimation.

### 4.2.2 Improved STT selection strategy using Gabor Filter

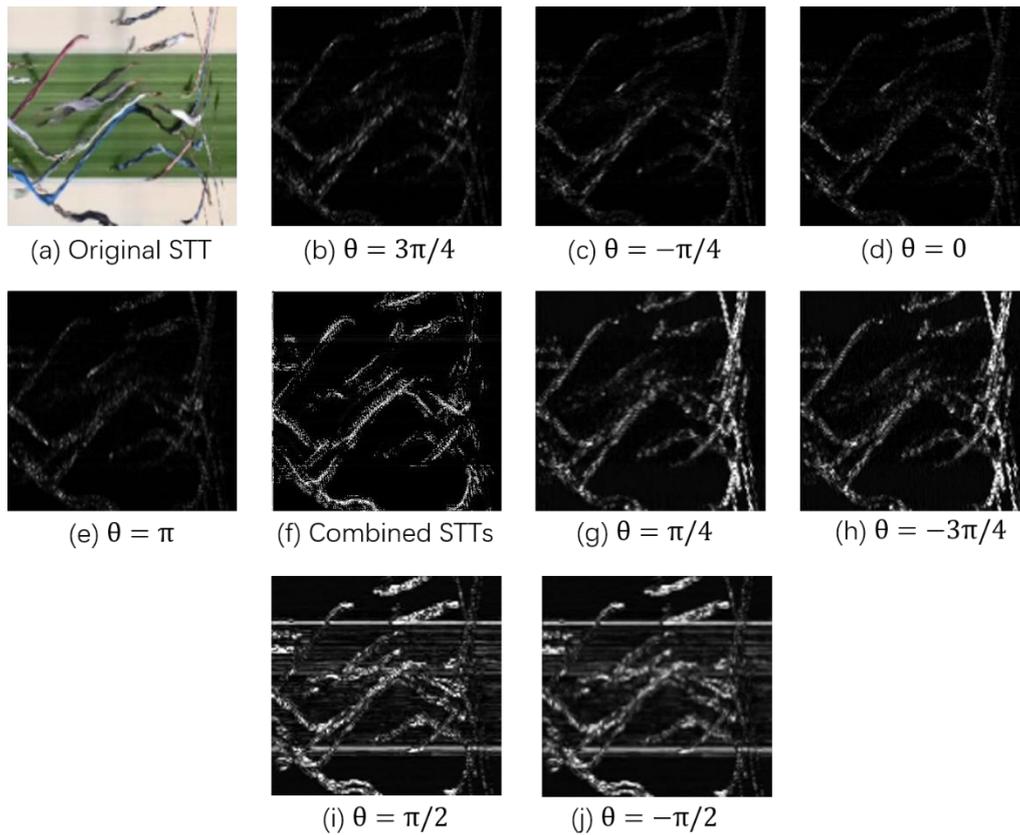
As explained in Figure 4-8, the static background only produces horizontal parallel lines. On the other hand, the motion trajectory in STT may be along any orientation. It can be recalled that the Gabor boundary detection approach uses Gabor kernel along various orientations to detect edges along all possible directions. Therefore, these horizontal parallel lines will also be detected as edges by Gabor kernel with  $\theta = \pi/2$  that can be removed from the kernel collection. With the new kernel settings, the motion trajectories in STT are extracted using the devised four-phases boundary detection

approach. The improved boundary detector can be considered as a background subtractor for the STT. According to the improved STT selection approach, the information entropy is calculated, and the STT with more motion patterns instead of background lines will be identified. The improved STT selection approach is illustrated in Figure 4-9.



**Figure 4-9.** The procedure of Improved STT selection approach

To verify the effectiveness with the new STT selection strategy, the background subtraction results are tested with and without the kernel function  $\theta = \pi/2$  for Gabor filtering. In the first implementation, the Gabor filter had been applied eight times along each direction, namely, N, S, W, E, SW, NW, SE, and NE. The filtered STT along each direction is illustrated in Figure 4-10. Figure 4-10(a) shows the original STT. Figure 4-10(b) to Figure 4-10(e) shows the results along W, E, S, N directions. Figure 4-10(g) to Figure 4-10(j) shows the results along SW, NW, SE, NE directions. After the eight results are obtained, they are integrated together as the final target STT for analysis. In this case, the parameters are set as follows: the wavelength is set to 2, the Standard Deviation is set to 0.45 (b-e), and 0.5 (g-j).



**Figure 4-10. Gabor filtering results along eight directions**

The experiments were carried out on 7 video clips from the UMN dataset (Cui *et al.*, [2011](#)). The selected STTs are annotated in the first column of Figure 4-11. Comparing to the results without background subtraction (see Figure 4-7), the identified STT of UMN2 contains more motion trajectories and less parallel lines. However, STTs of UMN3, UMN4 and UMN5 are still not satisfactory.

In the second implementation, the kernel function with  $\theta = \pi/2$  and  $\theta = -\pi/2$  are removed from the kernel collection. Therefore, Gabor background subtraction will be applied and the final integration is accumulated along six directions. The results provide a significant improvement. All 7 identified STTs contain detailed motion textures that are ideal for the follow up crowd behaviour analysis.

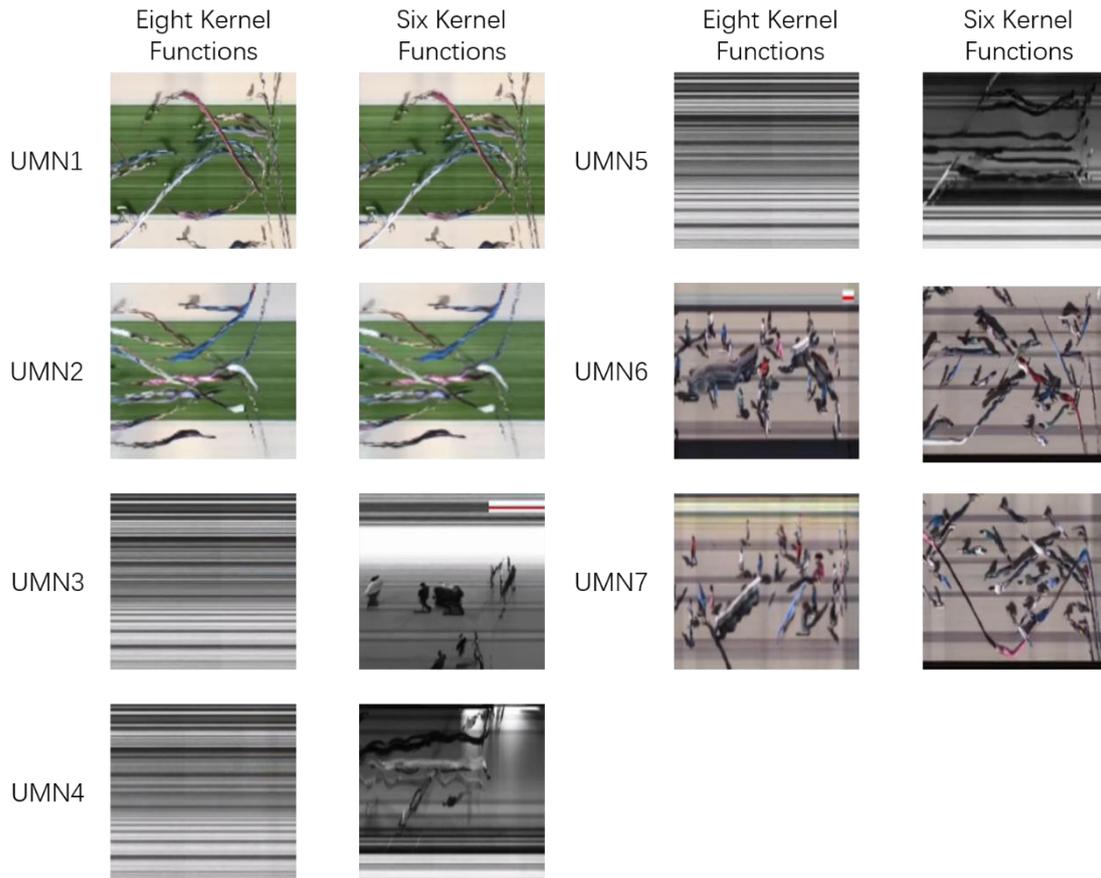


Figure 4-11. Comparison Between Results with eight and six kernel functions

### 4.2.3 Computational Efficiency

The devised texture selection technique is capable of significantly reducing the computational time on pattern extraction process such as the optical flow. According to the nature of flow-based pattern extraction approach, every pixel of each frame will be calculated. The maximum number of pixels can be as large as  $w * h * t$ , where  $w$  denotes the width of current frame,  $h$  denotes the height, and  $t$  denotes the total frame numbers. The possible computational complexity of the approach can reach  $O(n^3)$ . In contrast, the proposed STT extraction approach only requires the collection of texture data from several sampling locations. Therefore, the possible number of pixels to be calculated will decrease to  $(N + 1)wt + (M + 1)ht$ , and the potential computational complexity will be  $O(n^2)$ . Since various behaviours will exhibit different unique patterns in the STT, if the STT with proper patterns is selected with the proposed technique, these patterns can be modelled as signature (feature vectors) for

behaviour analysis in the follow-on process. Different from the change detection of panic behaviour based on the optical flow, the recognition of different abnormal behaviours can only be achieved with the classification of modelled pattern signatures.

#### 4.2.4 Exploiting the STT for Panic Detection

Since the extracted STT using proposed approach contains rich motion information, it is ideal for feature modelling and behaviour recognition. In this section, a sample application of STT for crowd panic dispersing detection is introduced to exhibit the STT's potent nature.

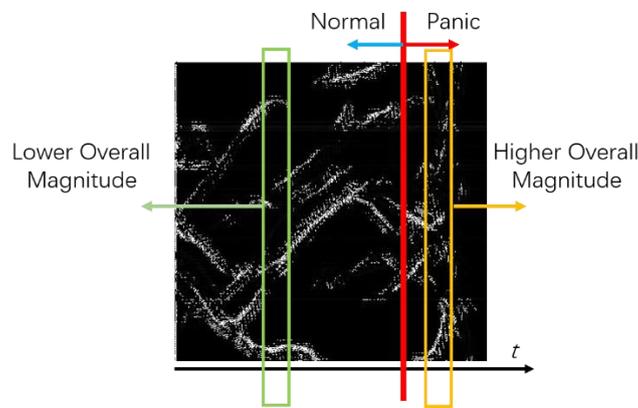


Figure 4-12. Motion Information in the STT

Figure 4-12 illustrates an extracted STT. Marked by the time-axis  $t$ , the STT consists of normal and panic states. In the normal state, pedestrians move with low velocity. The corresponding texture in STT is sparse, and the slope of trajectory is low. When the panic occurs, pedestrians begin to move with high velocity. The texture becomes condense, and the slope becomes steeper. A panic dispersing detection model is established. For pixels belong to the same time  $t$ , their Grey scale values are accumulated. If the difference between the average value and the summation in any time is greater than a threshold, this frame will be considered abnormal. The pseudo code of this algorithm can be expressed as in Listing 4-2.

---

```

I ← STT                                % Obtain STT
I ← ToGrayScale(I)                      % Transform to Gray Scale Image
T ← Constant                             % Assign a constant value to the threshold
(h, t) = size(I)                         % Obtain the height and time length of STT
A = 0
For i = 1: t'                             % For the first several frames
    Mi = 0
    For j = 1: h
        Mi = Mi + I(i, j)             % Calculate the summation of Gray Level
    End
    A = A + Mi
End
A = A/t'                                 % Calculate the average of first several frames
For i = t': t                             % For the rest frames
    Mi = 0
    For j = 1: h
        Mi = Mi + I(i, j)             % Calculate the summation of Gray Level
    End
    IF Mi − A > T                         % If the difference is larger than threshold
        Return Abnormal                     % Return anomaly
    End
End

```

---

**Listing 4-2. Pseudo Code of Panic Detection using STT**

The proposed approach is applied on 6 videos of UMN containing panic dispersing behaviours. As illustrated in Figure 4-13, the first column shows the extracted STT. The second column shows the magnitude map filtered by the six-directional Gabor. The third column shows the trend of magnitude along  $t$ . For each  $t$ , the motion magnitudes are accumulated. When motion at  $t$  is more drastic, the total magnitude will be larger. The color-bars under the magnitude trend indicate the detection results. The horizontal bar corresponds the  $t$  of the magnitude trend. The Grey bar implies the first 100 frames of training phase. The black bar implies the normal state, and the white bar implies the detected anomalies. Comparing to the ground truth, all panic dispersing behaviours are successfully detected.

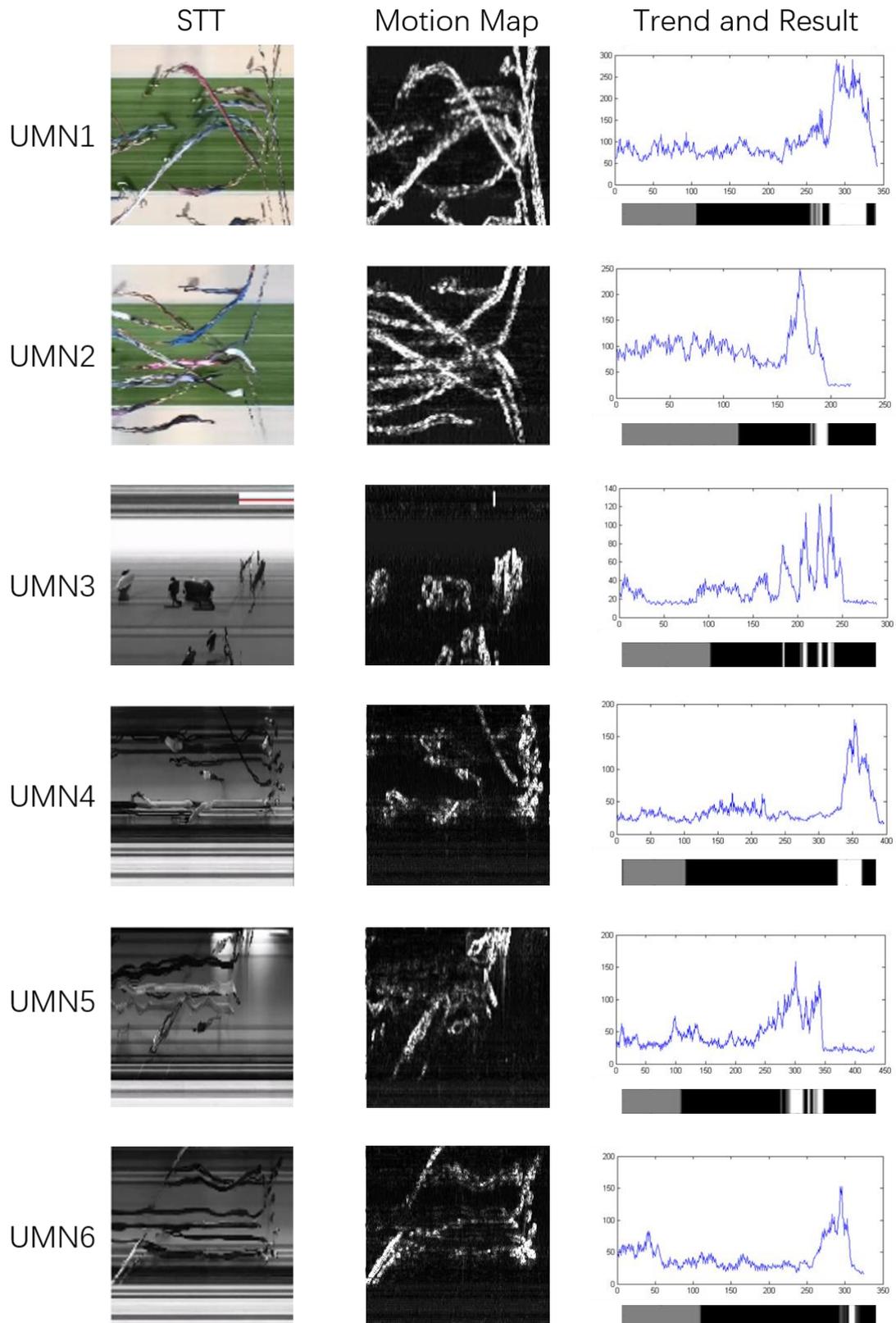


Figure 4-13. Crowd Panic Dispersing Detection Result

### 4.3 Chapter Summary

In this chapter, an effective spatial-temporal texture extraction technique based on information entropy and Gabor background subtraction is proposed in order to support the feature extraction and modelling process of the crowd analysis framework. The concept of information entropy is firstly introduced and proved to have the capability of represent the quantity of information within the image. Therefore, the information entropy is exploited for the selection of STT with most motion information. In order to get rid of the impact of static background lines within STT which could potentially generate high entropy, a six-directional Gabor filter background subtraction technique is devised to achieve the background removal on the STT. The preliminary experiment indicates the proposed approach has a good performance on STT selection. And the selected STT will be utilized for pattern extraction and modelling in next chapter.

## **Chapter 5. Crowd Behaviours Classification and Abnormality Detection**

As highlighted in last chapter, valid STT contains abundant crowd motion information and can be further processed for pattern modelling and classification. In this chapter, the Grey-Level Co-Occurrence Matrix (GLCM) texture patterns for STT are investigated. It is then modelled into crowd behaviour descriptors for classification.

The contents of this chapter are distributed as follows: Section 5.1 provides a brief overview of texture patterns, and covers the extraction approach of GLCM. Section 5.2 introduces the classifiers for behaviour recognition, and the rationale for adopting SVM. Section 5.3 explains the orderliness, descriptive and contrast features modelled from the primitive GLCM, and finally, the modelling and classification of the behaviour descriptors. Section 5.4 provides a case study on panic dispersion analysis.

## 5.1 Image Texture Patterns

The texture is a visual pattern revealing the homogeneity within an image. It describes the gradual or periodic change of the object's surface structure. The three key patterns of textures are: the repeating local sequence; the non-random distribution; and, the uniformity within a texture region. Different from the color and Grey level patterns, the local texture information is described by the Grey level distribution of pixel and its neighbors. The global texture information is the repetitiveness of local texture information.

Since the texture describes the surface property of corresponding objects, high-level semantic contents can be represented using only textures. Different from the color pattern, a piece of texture isn't based on the single pixel, it involves the statistical calculation of multiple pixels within a region. Using texture can effectively alleviate pattern matching problems caused by the regional deviation.

Using texture pattern is an effective approach when indexing images with large differences, for example, feature density. However, if the difference is not significant enough, normal texture patterns may fail to accurately reveal the nature of the observed targets. For example, the reflection on a specular surface will inundate effects from underlying textures of the object. In summary:

The advantages of using texture patterns.

- The statistical calculation of all pixels within the region instead of single pixel.
- Rotational invariance.

- Higher resistance of Noise.

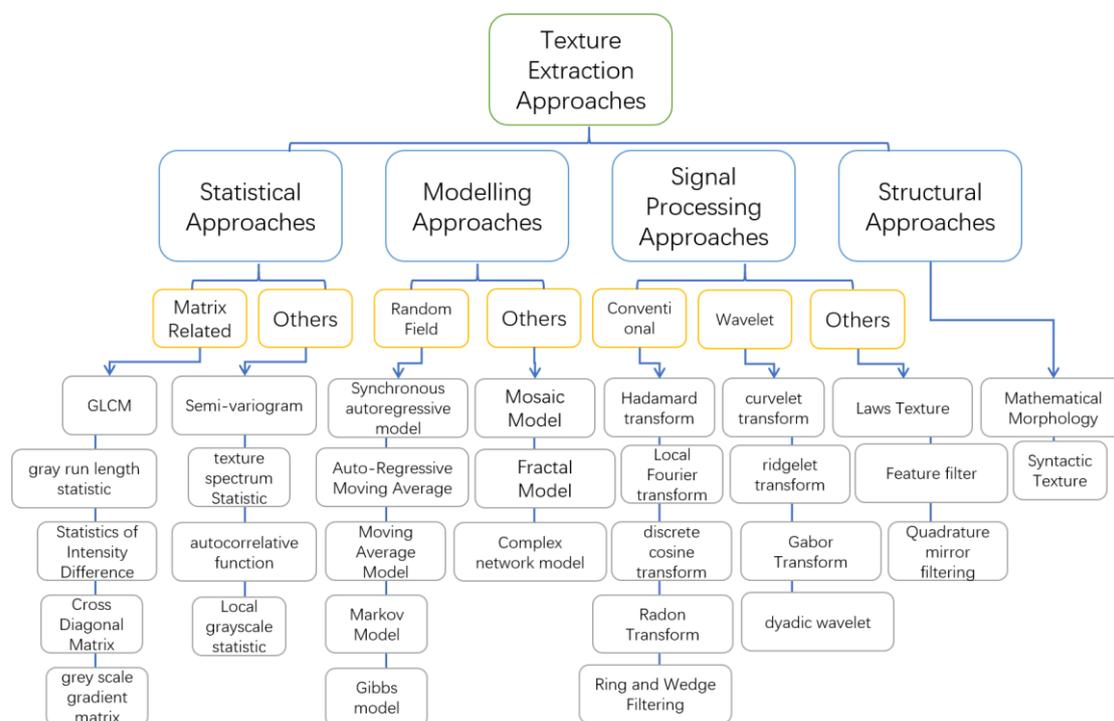
The disadvantages of using texture patterns.

- The resolution of image significantly affects the extracted texture.
- The brightness and reflection affect the extracted texture.
- The obtained texture from 2-D image doesn't always match the 3-D object's actual texture.

### **5.1.1 Taxonomy of Texture Pattern Extraction Approaches**

In order to extract the texture, a filtering window will be sliding through the image, pixels within the window are exploited for the texture calculation. However, the selection of window size is usually a dilemma. Texture is a regional concept, which is expressed by the spatial uniformity. The larger the window is, the easier to detect its uniformity. The change of uniformity indicates the boundary of different textures. Therefore, a smaller window can achieve the more accurate detection of texture boundary. In conclusion, if the size is too small, the inaccurate extraction occurs within the right texture. If the size is too large, the inaccurate extraction occurs between the boundary of textures.

Various texture extraction approaches have been examined. These approaches can be divided into four main categories as illustrated in Figure 5-1. These four categories consist of statistical, modelling, signal processing and structural approaches.



**Figure 5-1. Taxonomy of Texture Extraction Approaches**

#### a) Statistical Approaches

These approaches focus on the first, second or higher order statistical patterns in the texture region based on pixel and its neighbor's Grey level. The typical approaches are as follows: GLCM and its derived patterns such as Energy, Entropy and Correlation are computed; texture pattern extracted from the energy spectrum function of the image; a semi-variogram reveals the randomness and structure of the image.

The advantages of statistical approaches are the low modelling difficulty and high adaptiveness. Disadvantages are the disjoint to human visual model, and the absence of the global information and high computational complexity.

#### b) Modelling Approaches

The modelling approaches assume the texture pattern can be expressed as a parameter driven distribution model. Therefore, the core issue of this approach is the estimation of parameters. Typical approaches include Markov Random Field (Li, [1994](#)), Gibbs Random Field (Howard and Haluk, [1987](#)) and Fractal Model (Alex, [1984](#)).

The merits of this type of approaches include the flexibility of the balancing

between local randomness and global regularity features. The shortcomings are stemmed from the higher difficult parameter estimation, for example, the slow execution of the iterative process in MRF.

#### c) Signal Processing Approaches

The signal processing approaches usually transform certain region of the image into the temporal and spectrum domains, and then to extract the relatively stable patterns for expressing the uniformity within the region. Typical approaches include Gabor transform, Tamura texture and Wavelet transform.

The advantage is that it analyses textures in a more precise manner. The classic wavelet transform matches the human vision model and visual habits, which helps the segmentation of texture image. However, these approaches prefer regular texture patterns than the complex ones and having relative low performance on nature images.

#### d) Structural Approaches

The structural approaches assume the texture is composed with the permutation rules of texture elements, quantity of them, and the spatial structure. The core issues of this approach are the intrinsic difficult in extracting elementary texture elements and the spatial structure. The typical approaches are Syntactic Texture and Mathematical Morphology.

### 5.1.2 Grey Level Co-occurrence Matrix

The GLCM is also known as the Grey Tone Spatial Dependency Matrix, which is introduced in the research of Haralick *et al.* (1973). The GLCM expresses the statistic distribution of the different Grey scale value levels within an image. For any pixel pairs  $(i, j)$  and  $(i + a, j + b)$ , the corresponding Grey levels are  $(f_1, f_2)$ . Assume the maximum Grey level is  $L$ , the combination of  $(f_1, f_2)$  will be  $L * L$ . For the entire image, frequency of each  $(f_1, f_2)$  combination is modelled as a matrix  $G$ . Next,  $G$  is normalized with the total number of combinations into  $P(f_1, f_2)$  as the GLCM. In real-life video, the extracted STT doesn't have the regular patterns. Therefore,  $G$  is often asymmetric. Since  $G$  only aggregates the relation along single direction. In order to

make  $G$  representing bi-directional relations, the transposing matrix  $G'$  is calculated. Then the summation  $S$  of  $G$  and  $G'$  is calculated. Next, the probability matrix  $P$  could be expressed as Equation 5-1.

$$P_{i,j} = \frac{S_{i,j}}{\sum_{i,j=0}^{N-1} S_{i,j}} \quad 5-1$$

Where  $i$  denotes the row index of matrix, and  $j$  denotes the column index. According to the definition, the probability matrix  $P$  contains two unique properties. 1) Size of  $P$  is determined by the number of  $L$ . For example, assuming the Grey scale values are divided into three levels, the total row and column number of  $P$  will be 3. Therefore, when the  $L$  becomes larger, size of  $P$  will increase, and the pattern of  $P$  is more explicit. On the other hand, if the  $L$  is too high, the distribution of  $P$  will be sparser. In this situation, descriptive capability of  $P$  will be greatly impacted. In practice, the range of the level number is usually set among 3 and 10. Furthermore, the proper  $L$  could reduce the time consumption. 2) Since  $P$  is symmetric, its diagonal where  $i - j = 0$  denotes the combinations without Grey level differences. When the matrix indices stay far from the diagonal where  $|i - j|$  is large, the Grey level differences become larger for the pixel pair. Therefore, more pixel pairs distributed in far side of the diagonal indicate the image has higher contrast value.

An example is illustrated in Figure 5-2(a) - an image with 7 by 7 pixels. As shown in Figure 5-2(b), the highest Grey level is 14 and the lowest Grey level is 2, total number of levels is 4. The 4 Grey levels are replaced with 0 to 3 as illustrated in Figure 5-2(c), Figure 5-2(d) and Figure 5-2(e). It also means the range of  $(f1, f2)$  is from 0 to 3. By setting the different  $a$  and  $b$ , and permutating the number of different  $(f1, f2)$  combinations, the  $G$  illustrated in Figure 5-2(f)-(g) is obtained.

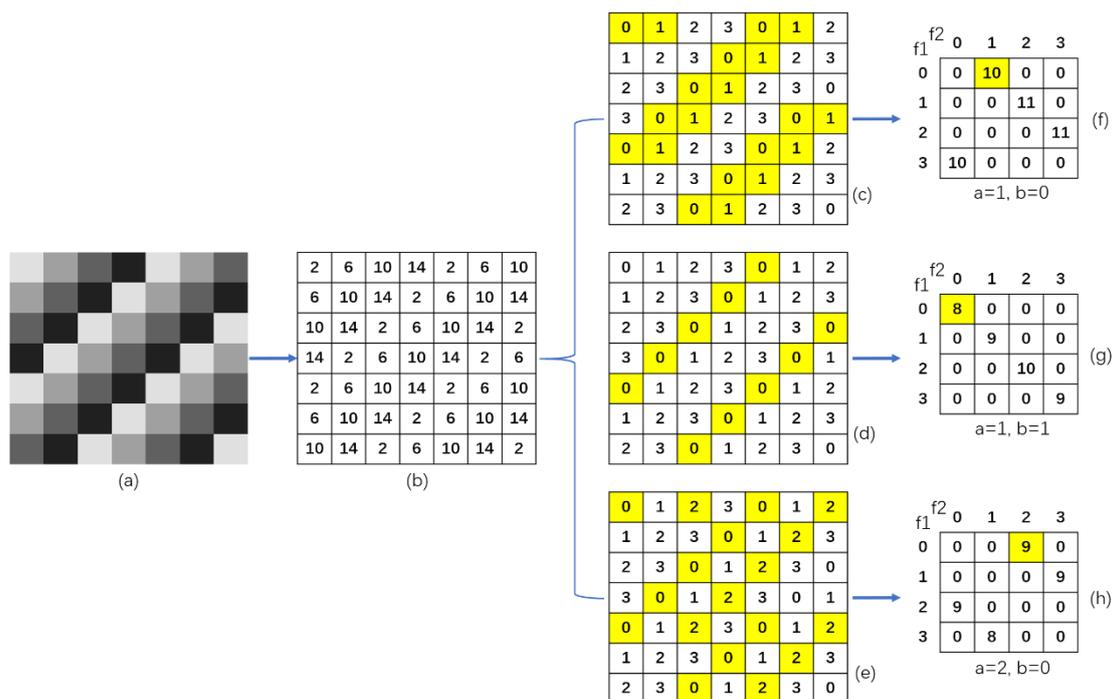


Figure 5-2. GLCM extraction

The labelled slots in Figure 5-2(c) indicates the pixels pairs with  $(f_1 = 0, f_2 = 1)$ , and  $(a = 1, b = 0)$ . The total number of this combination is 10. Therefore, the corresponding slot of GLCM is set to 10 in Figure 5-2(f). Similarly, labelled slots in Figure 5-2(d) indicates the combination of  $(f_1 = 0, f_2 = 0)$  and  $(a = 1, b = 1)$ , and labelled slots in Figure 5-2(e) indicates the combination of  $(f_1 = 0, f_2 = 2)$  and  $(a = 2, b = 0)$ . The GLCM in Figure 5-2(f) indicates the  $(f_1 = 0, f_2 = 1)$ ,  $(f_1 = 1, f_2 = 2)$ ,  $(f_1 = 2, f_2 = 3)$  and  $(f_1 = 3, f_2 = 0)$  have higher frequency. Therefore, the corresponding image has significant texture pattern from left-bottom to right-top.

The different values of  $(a, b)$  will determine the different GLCM. The selection of  $(a, b)$  should depend on the distribution of texture's pattern. For the narrow texture, the value could be pairs such as  $(0,1)$  or  $(1,1)$ . For the slow changing textures, the small  $(a, b)$  will derive the larger values on the diagonal. On the other hand, the fast-changing texture will make the smaller values on diagonal, and larger values on far-side. The pseudo code to calculate the GLCM is illustrated in Listing 5-1.

---

```

I ← STT                                % Obtain STT
I ← ToGrayScale(I)                    % Transform to Gray Scale Image
L ← Constant                            % Set the number of gray levels
(a, b)                                    % Set the calculation direction
G(L, L)                                  % Initialize the GLCM with size L by L
(h, w) = size(I)                        % Obtain the height and time length of STT
For i = 1: h                            % For each pixel
    For j = 1: w
        m = I(i, j) * L/256                % Normalize the f1 into the L scale
        n = I(i + a, j + b) * L/256        % Normalized the f2 into the L scale
        G(m, n) = G(m, n) + 1            % Increase the (f1, f2) by 1
    End
End

S = G + G'                              % Summation of G and its transpose
P(L, L)                                  % Initialize the Probability Matrix
T = 0                                      % The accumulation of S
For i = 1: L                              % For each pixel
    For j = 1: L
        T = T + S(i, j)                    % Obtain the accumulation of S
    End
End

For i = 1: L                              % For each pixel
    For j = 1: L
        P(i, j) = S(i, j)/T                % Calculate the Probability Matrix
    End
End

```

---

Listing 5-1. pseudo code of GLCM calculation

## 5.2 Machine Learning Classifiers

According to the proposed framework of crowd behaviour detection, once the texture patterns are extracted and modelled into descriptor, the classifier will be utilized to determine whether the descriptor is abnormal. The capability of classifiers greatly

affects the detection result. Therefore, the appropriate selection of classifier is crucial to the entire operation. In this section, conventional machine learning classifiers are introduced, including K-Nearest Neighbors, Support Vector Machine and Back Propagation Neural Network. And the reason of choosing SVM as the classifier for the crowd behaviour detection is explained.

### 5.2.1 K-Nearest Neighbours

Assume a training sample set exists, all data is labelled with different properties. When the test data is input, compare every property between test and training data, and find out the nearest samples. Next, the first k samples are used for analysis. The type exists most in these samples will be selected as the type of test data. The process of KNN is illustrated as Figure 5-3. If the k is set to 3, the type of circle will be classified as Triangle. Since there are 2 triangles in the 3 nearest neighbors. Similarly, If the k is set to 5, the type of circle will be classified as Rectangle.

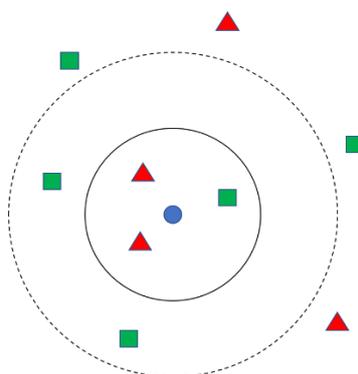


Figure 5-3. K-Nearest Neighbor Classification

The advantages of KNN include the insensitive to rare value, low difficulty to achieve, and adaptiveness to multi-model classification. The primary disadvantage of KNN is the high computational and spatial complexity.

### 5.2.2 Support Vector Machine

The Support Vector Machine is a supervised generalized linear classifier based on binary classification. Its decision boundary is the maximum-margin hyperplane of learning sample's solution. The SVM is firstly proposed in 1963, and various derived

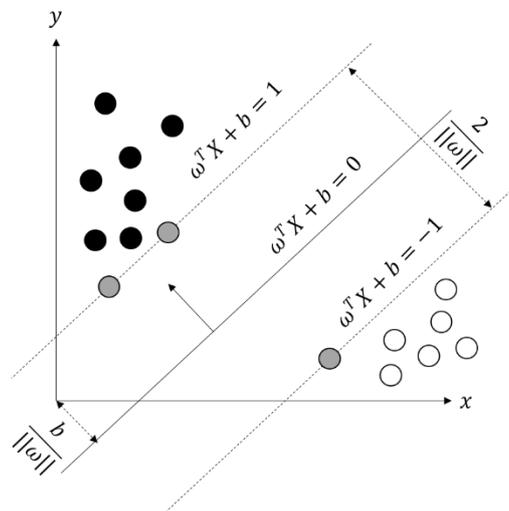
enhancements are developed, including Multi-class SVM, Least-Square SVM, Support Vector Regression, Support Vector Clustering and Semi-Supervised SVM. The SVM is widely exploited in face recognition, text categorization and pattern recognition. The conventional SVM consists of two approaches, which are Linear and Kernel approaches.

#### a) Linear Approach

Given the input data  $X = \{X_1, \dots, X_N\}$  and class set  $y = \{y_1, \dots, y_N\}$ , where each data consists of multiple features as a feature space  $X_i = [x_1, \dots, x_n] \in \chi$ . The binary parameter  $y \in \{-1, 1\}$  is used to label the positive and negative class. If a hyperplane  $\omega^T X + b = 0$  of decision boundary in the feature space exists and divides the data into positive and negative, and the distance between the hyperplane and any sample is larger than 1  $y_i(\omega^T X_i + b) \geq 1$ , then these data is considered linearly separable. The parameters  $\omega, b$  are the normal vector and intercept of the hyperplane. Two parallel hyperplanes  $\omega^T X + b = \pm 1$  are modelled as the interval boundary to classify the samples, as expressed in Equation 5-2.

$$\begin{aligned} \omega^T X_i + b - 1 &\geq +1, & \text{if } y_i = +1 \\ \omega^T X_i + b + 1 &\leq -1, & \text{if } y_i = -1 \end{aligned} \quad 5-2$$

Data larger than the upper boundary is positive, and data smaller than the lower boundary is negative. The distance  $d = \frac{2}{\|\omega\|}$  between two boundaries are defined as Margin. The samples fall right on the boundaries are defined as Support Vectors. As illustrated in Figure 5-4, the Grey circles indicate the support vectors. The dash lines indicate the interval boundaries. The full line indicates the decision boundary.



**Figure 5-4. Support Vectors, Decision Boundary and Interval Boundary in SVM**

When a classification problem is not linearly separable, using the hyperplane will cause the classify losing. In this situation, some support vectors fall into the interval boundary or the wrong side of decision boundary. The Loss Function could quantify the classify losing. The 0-1 loss function could be defined as Equation 5-3.

$$L(p) = \begin{cases} 0 & p < 0 \\ 1 & p \geq 0 \end{cases} \quad 5-3$$

Since the 0-1 loss function isn't continuous, which isn't appropriate for optimization. Regular solution is exploiting the surrogate loss, including Hinge Loss, Logistic Loss and Exponential Loss, expressed as Equation 5-4. The SVM uses hinge loss function.

$$\begin{aligned} \text{hinge: } L(p) &= \max(0, 1 - p) \\ \text{logistic: } L(p) &= \log[1 + \exp(-p)] \\ \text{exponential: } L(p) &= \exp(-p) \end{aligned} \quad 5-4$$

#### b) Kernel Approach

Some linearly separable problems could be non-linearly separable. A hypersurface exists in the feature space to divide the positive and negative samples. By using the non-linear function, the non-linearly separable problem could be mapped into higher-dimensional Hilbert Space  $H$  from its original feature space to transform the problem into linear separable one. The hypersurface as the decision boundary could be expressed as Equation 5-5.

$$\omega^T \phi(X) + b = 0 \quad 5-5$$

Where  $\phi: \chi \rightarrow H$  is the mapping function. Since the mapping function is complex non-linear function, therefore, the kernel function could be exploited to simplify the computation. The conventional kernel functions are listed as follows. When the  $n$  is 1, the Polynomial Kernel becomes linear, the corresponding classifier becomes linear.

- Polynomial Kernel:  $k(X_1, X_2) = (X_1^T X_2)^n$
- RBF Kernel:  $k(X_1, X_2) = \exp(-\frac{\|X_1 - X_2\|^2}{2\sigma^2})$
- Laplacian Kernel:  $k(X_1, X_2) = \exp(-\frac{\|X_1 - X_2\|}{2\sigma^2})$

- Sigmoid Kernel:  $k(X_1, X_2) = \tanh[a(X_1^T X_2) - b]$ ,  $a, b > 0$

### 5.2.3 Back Propagation Neural Network

Back Propagation (BP) Neural Network is an effective multi-level neural network learning approach. In BP network, the signal is transferred forward and the deviation is transferred backward. By continuously adjusting the weight of network, the actual output will be gradually close to the desired output. The structure of a typical BP network is illustrated in Figure 5-5.

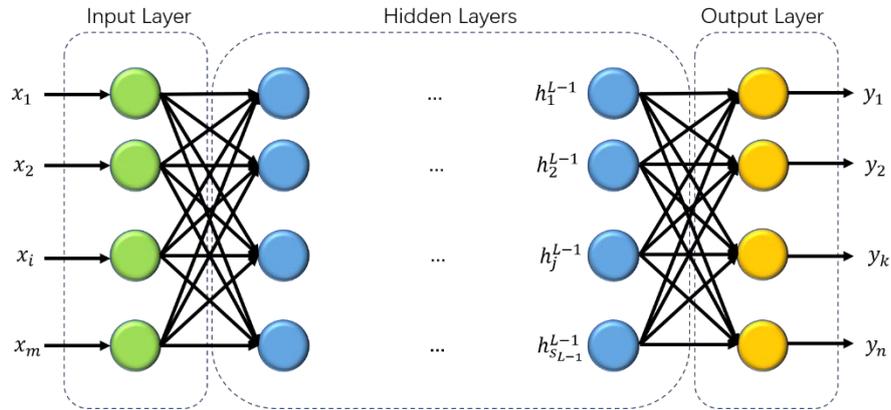


Figure 5-5. Structure of BP Neural Network

The BP network consists of  $L$  neural layers. The first layer is input layer, the  $L$ th layer is output layer. The 2<sup>nd</sup> to  $L - 1$  layers are hidden layers. Where the input vector is

$$\vec{x} = [x_1 \ x_2 \ \dots \ x_i \ \dots \ x_m], i = 1, 2, \dots, m \quad 5-6$$

The output vector is

$$\vec{y} = [y_1 \ y_2 \ \dots \ y_k \ \dots \ y_n], k = 1, 2, \dots, n \quad 5-7$$

The output of  $l$ th hidden layer is

$$h^{(l)} = [h_1^{(l)} \ h_2^{(l)} \ \dots \ h_j^{(l)} \ \dots \ h_{s_l}^{(l)}], j = 1, 2, \dots, s_l \quad 5-8$$

Where  $s_l$  is the number of neural in  $l$ th layer.

Assuming  $W_{ij}^{(l)}$  is the weight of connection between the  $j$ th neural in  $l - 1$  layer and  $i$ th neural in  $l$ th layer.  $b_i^{(l)}$  is the offset of the  $i$ th neural in  $l$ th layer. Thus

$$\begin{aligned}
 h_i^{(l)} &= f(\text{net}_i^{(l)}) \\
 \text{net}_i^{(l)} &= \sum_{j=1}^{s_{l-1}} W_{ij}^{(l)} h_j^{(l-1)} + b_i^{(l)}
 \end{aligned}
 \tag{5-9}$$

Where  $\text{net}_i^{(l)}$  is the input of the  $i$ th neural in  $l$ th layer.  $f(\cdot)$  is the activation function of the neural. In most cases, the activation function will be a non-linear. The following two activation functions are usually adapted in BP network. The first one is the Sigmoid function, and the second one is the Hyperbolic Tangent function.

$$\begin{aligned}
 f(x) &= \frac{1}{1 + e^{-x}} \\
 f(x) &= \frac{1 - e^{-x}}{1 + e^{-x}}
 \end{aligned}
 \tag{5-10}$$

The BP approach could be expressed as following processes.

- For all layers  $2 \leq l \leq L$ , set  $\Delta W^{(l)} = 0, \Delta b^{(l)} = 0$ , where  $\Delta W^{(l)}$  is a zero matrix and  $\Delta b^{(l)}$  is a zero vector.
- For each  $i$  between 1 to  $m$ 
  - Calculating the gradient matrix of neural weight  $\nabla W^{(l)}$  and offset  $\nabla b^{(l)}$  using back-propagation algorithm.
  - Calculating  $\Delta W^{(l)} = \nabla W^{(l)}(i)$ .
  - Calculating  $\Delta b^{(l)} = \nabla b^{(l)}(i)$ .
- Updating the weight and offset
  - Calculating  $W^{(l)} = W^{(l)} + \frac{1}{m} \Delta W^{(l)}$ .
  - Calculating  $b^{(l)} = b^{(l)} + \frac{1}{m} \Delta b^{(l)}$ .

The advantages of BP network are its non-linear reflection capability and flexible network structure. The number of hidden layers and neural could be customized according to the actual situation. The primary disadvantages are the relatively slow training speed and the traps of falling into the local minimum.

In this research, the SVM is more appropriate than the BP network on the behaviour classification using STT. The main reason is the requirement of the large training set in neural network. However, the number of crowd behaviour video dataset is small and

not labelled. Compared to the BP network, the training data requirement for SVM is much lower. Therefore, the SVM is selected as the classifier.

### 5.3 Classification using GLCM

In order to exploit the STT, an approach of modelling the texture signature based on the GLCM patterns is devised for the recognition of abnormal crowd behaviour. Firstly, GLCM will be calculated. Next, these GLCM patterns will be further modelled into a signature/descriptor. The modelled signature will be utilized for training and classification of behaviours in STT. In the experiment, the results exhibit high accuracy on the detection of abnormal crowd behaviours such as panic dispersing and congestion. The modelled STT signature is proven to be an efficient descriptor for the automatic crowd analysis in practice.

#### 5.3.1 Modelling Features From GLCM

The GLCM only represents the Grey level distribution of the texture, in order to more explicitly represent the motion information in the STT, the GLCM is further modelled into different features. As illustrated in Figure 5-6, once the target STT is extracted, the corresponding GLCM will be calculated. Next, the Orderliness features, Descriptive features and Contrast features will be modelled from the raw GLCM. Then, a signature will be modelled from these features. Finally, the signature will be used for the classification of crowd behaviours.

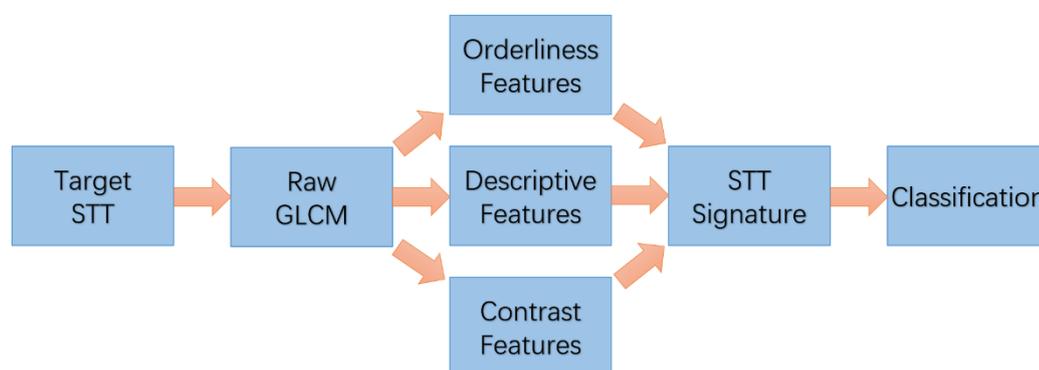


Figure 5-6. Structure of the Signature Modelling approach

As illustrated in the third Step of Figure 5-6, the approaches of modelling the three features from the GLCM matrix  $P$  are introduced in next section.

### 5.3.2 Contrast Features of GLCM

By definition, the Contrast type features describe the acuteness of the Grey level changing between neighboring pixel pairs. The Contrast type consists of four features, which are contrast, dissimilarity, homogeneity and similarity. For contrast, when the pixel pair at row  $i$  and column  $j$  is far from the diagonal of GLCM probability matrix  $P$ , the global contrast value will be larger. The contrast value  $CON$  can be obtained from  $P$  with Equation 5-11.

$$CON = \sum_{i,j=0}^{N-1} P_{i,j} (i - j)^2 \quad 5-11$$

The Dissimilarity feature is similar to the contrast. It also represents how drastic the Grey level changes in the texture by replacing the exponential weight with the linear weight. The dissimilarity could be denoted as  $DIS$  and shown as Equation 5-12.

$$DIS = \sum_{i,j=0}^{N-1} P_{i,j} |i - j| \quad 5-12$$

The Homogeneity represents the consistency of the STT, which is also known as Inverse Different Moment (IDM). The changing trend of Homogeneity is exactly opposite to the Contrast. When the change of Grey level is less drastic, the value of Homogeneity will be larger. Equation 5-13 indicates the expression of Homogeneity as  $HOM$ .

$$HOM = \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1 + (i - j)^2} \quad 5-13$$

Same to the relation between Contrast and Homogeneity, the Similarity feature is opposite to the Dissimilarity according to Equation 5-14.

$$SIM = \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1 + |i - j|} \quad 5-14$$

The Table 5-1 illustrates the comparison between the Contrast type features of various STT patches. These patches are collected from different positions and states in

the same STT. In the experiment, the window size of raw GLCM is set to 50 by 50. The sampling direction is horizontal, and the sampling step is set to 1. The level number of Grey level is set to 8. The results indicate the patch in Table 5-1(a) obtains higher homogeneity and similarity, but lower contrast and dissimilarity values than Table 5-1(d), because it contains motion information of panic behaviour.

### 5.3.3 Orderliness Features of GLCM

The Orderliness type features describe whether the changing of the Grey level between neighboring pixel pairs is regular. This type consists of three features, including Angular Second Moment (ASM), Energy and Entropy. The ASM is usually exploited for the evaluation of rotational acceleration. The ASM is expressed as Equation 5-15.

$$ASM = \sum_{i,j=0}^{N-1} P_{i,j}^2 \quad 5-15$$

The Energy feature reveals the similar nature of texture to ASM, except the Energy (ENR) is the square root of ASM's value shown as Equation 5-16. The Energy feature could be utilized in the research of fingerprint and botany.

$$ENR = \sqrt{ASM_{i,j}} \quad 5-16$$

The Entropy feature is opposite to Energy. Instead of regularity, this feature reveals the level of irregularity of the pixel distribution. If the motion in the STT is at the chaotic state, the entropy feature would be higher. The Entropy pattern ENT is shown as Equation 5-17.

$$ENT = \sum_{i,j=0}^{N-1} P_{i,j}(-\ln P_{i,j}) \quad 5-17$$

The Table 5-1 also illustrates the obtained values of Orderliness type features. The distribution of different features follows their definitions. For example, the Entropy value of patch in Table 5-1(a) is smaller than the one in Table 5-1(d), since the previous one contains more motion information.

### 5.3.4 Descriptive Statistical Features of GLCM

The Descriptive Statistical type features are obtained by calculating the statistical information of the GLCM. This type consists of three features including Mean, Variance and Correlation. The Mean feature could be expressed as Equation 5-18. The  $\mu_i$  and  $\mu_j$  indicate the Mean value along the row and column respectively. Since the matrix  $P$  is symmetric, the mean values along these directions equals to each other.

$$\mu_i = \sum_{i,j=0}^{N-1} iP_{i,j} \quad 5-18 \text{ (a)}$$

$$\mu_j = \sum_{i,j=0}^{N-1} jP_{i,j} \quad 5-18 \text{ (b)}$$

Similarly, the Variance of  $P$  is obtained with corresponding Mean  $\mu$ , and marked as  $\sigma^2$ . And the value of Deviation is the square-root of the Variance, and marked as  $\sigma$ . Both of these patterns could be expressed as Equation 5-19.

$$\sigma_i^2 = \sum_{i,j=0}^{N-1} P_{i,j}(i - \mu_i)^2 \quad 5-19 \text{ (a)}$$

$$\sigma_j^2 = \sum_{i,j=0}^{N-1} P_{i,j}(j - \mu_j)^2 \quad 5-19 \text{ (b)}$$

As the last feature, the Correlation COR could be obtained with the previously calculated Mean and Variance, which could be expressed as Equation 5-20.

$$COR = \sum_{i,j=0}^{N-1} P_{i,j} \left[ \frac{(i - \mu_i)(j - \mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} \right] \quad 5-20$$

### 5.3.5 GLCM Signature Modelling

According to the classification procedure illustrated in Figure 5-6, once the above introduced features are obtained from the probability matrix, they will be modelled into a signature for the behaviour classification. In this section, the performance of these features is tested and evaluated. Four key features will be selected and modelled in order to derive the final signature.

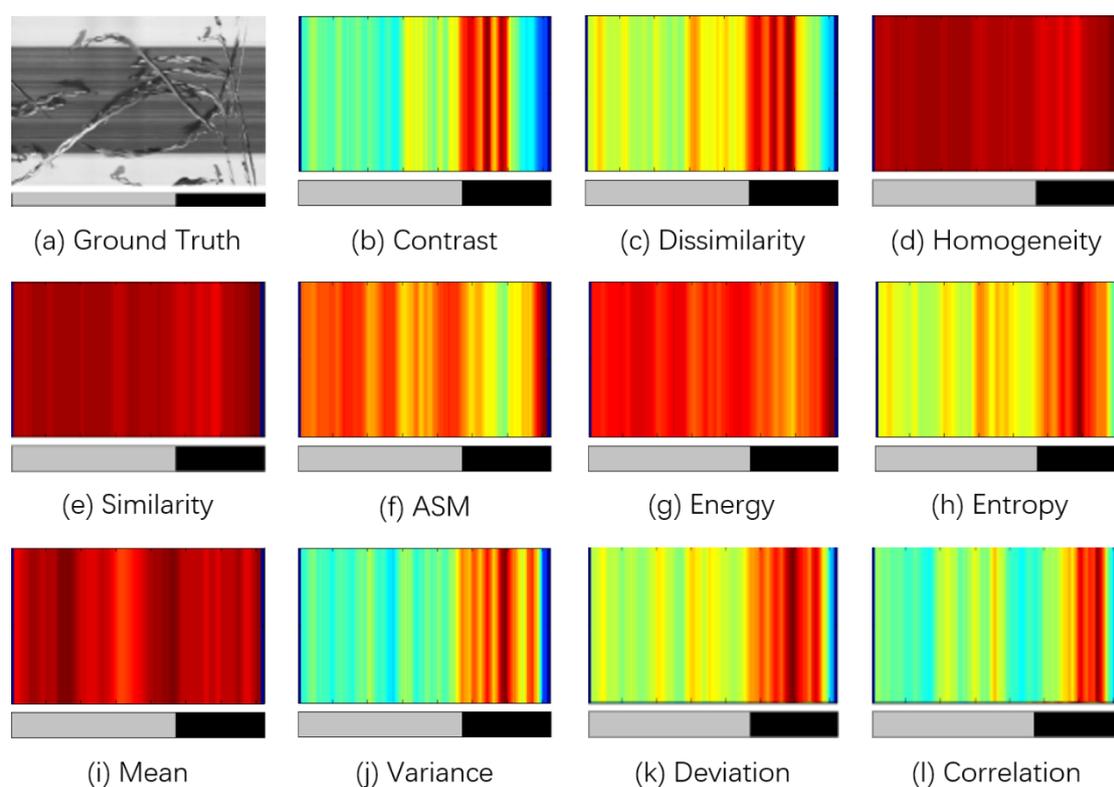
The first row of Table 5-1 illustrates six STT patches. These patches are extracted

from a single STT at different positions. The patch(a) to patch(c) are extracted from the normal STT. And the patch(d) to patch(f) are from the abnormal STT. Next, the values of Contrast, Orderliness and Descriptive Statistical features are calculated. Then, the trend of these values will be inspected. Firstly, patches with the abnormal behaviours tend to have higher values on Contrast, Entropy and Variance. For example, the Entropy values of patches from (a) to (c) are lower than those from (d) to (f). Secondly, STT patches with abnormal behaviours will have lower feature values than the normal ones, such as ASM. Furthermore, comparing to the changing trend of other features, Contrast, ASM, Entropy and Variance exhibit the most distinguished turbulence in value from normal to abnormal. Therefore, these four patterns are considered the most symbolic features to describe the behaviour in STT.

In the following experiment, the global changing trend of these patterns on a STT is probed. As illustrated in Figure 5-7(a), the STT extracted from the first sequence of UMN dataset is used for the feature modelling. Videos from the UMN consist of two stages including normal state and panic dispersing. Therefore, the STT shown in Figure 5-7(a) can be divided into two sections. A color bar at the bottom of figure labels the ground truth of STT. The section located in the Grey bar represents the normal state, and the one located in the black bar represents the panic state. The labelled ground truth generally matches the abnormal trajectories in STT. Therefore, the patterns extracted from the abnormal section of STT should also be able to represent the labelled ground truth. Since the STT is cut along the time axis, the width of STT equals to the length of video. By accumulating the pixels in each column, the global changing trend of these GLCM features along time could be observed.

Changing trends of four different Contrast features' magnitude are illustrated in Figure 5-7(b-e). When the abnormal behaviour occurs, the pattern values of Contrast and Dissimilarity indicate a significant surging. On the other hand, the pattern values of Similarity and Homogeneity will have a relatively smooth change. Figure 5-7(f-h) list the changing trends of Orderliness type features extracted from the STT in Figure 5-7(a). When the panic dispersing occurs, the irregularity describing features such as

Entropy increase quickly. On the contrary, the overall value of ASM suffers from a major decline. The trend of Energy isn't sensitive to the abnormality. The changing trends of Descriptive Statistical type features are illustrated in Figure 5-7(i-l). When the abnormal behaviour occurs, the trend of Mean pattern is relatively smooth with the Variance, Standard Deviation and Correlation. Based on the assumption that the more sensitive the feature reacts to different state, the better performance for the behavioural analysis, four features are selected as the key patterns for the signature including CON, ASM, ENT and VAR. Despite the Dissimilarity is also sensitive, it is still given up since it has the same trend with CON. Same reason to the features such as Standard Deviation and Correlation. The modelled signature is a combination of the selected patterns, which is shown as Equation 5-21. The corresponding experiments to evaluate the performance of proposed signature will be implemented on different types of crowd behaviours in chapter 7.



**Figure 5-7. Trends of GLCM patterns along time**

$$SIG = [CON, ASM, ENT, VAR]$$

5-21

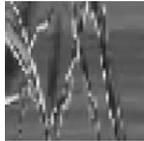
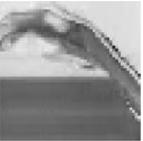
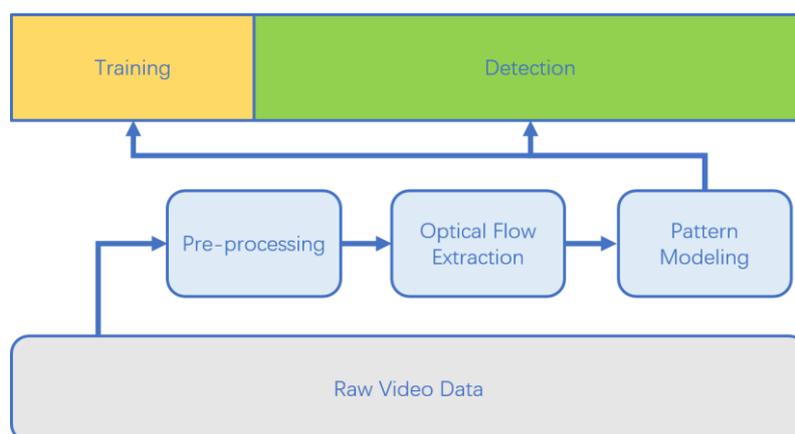
	Patch (a)	Patch (b)	Patch (c)	Patch (d)	Patch (e)	Patch (f)
						
<b>Contrast</b>	0.2437	0.3237	0.2669	0.6853	0.5735	0.6473
<b>Dissimila</b>	0.1922	0.2110	0.1935	0.3947	0.3645	0.4278
<b>Homogen</b>	0.9085	0.9049	0.9103	0.8302	0.8379	0.8078
<b>Similarity</b>	0.9101	0.9094	0.9132	0.8405	0.8459	0.8174
<b>ASM</b>	0.3538	0.2134	0.4124	0.1853	0.2062	0.1767
<b>Energy</b>	0.5948	0.4619	0.6422	0.4304	0.4541	0.4203
<b>Entropy</b>	1.2977	2.0858	1.5294	2.3325	2.1747	2.3599
<b>Mean</b>	2.4933	4.7598	2.7865	4.0635	2.6410	2.8265
<b>Variance</b>	0.3859	3.7115	0.6728	2.5150	1.0509	2.8265
<b>Deviation</b>	0.6212	1.9265	0.8202	1.5859	1.0251	1.0729
<b>Correlat</b>	0.6843	0.9564	0.8016	0.8638	0.7272	0.7188
<b>Normal</b>	Yes	Yes	Yes	No	No	No

Table 5-1. Comparison between texture patterns of Spatio-Temporal Texture patches

## 5.4 Case Study: Real-time Change Detection

The previously proposed crowd behaviour detection approach using GLCM and SVM involves the STT extraction, GLCM calculation, Signature modelling and classification. The procedure takes long training process and large quantity of training data. Despite achieving high accuracy, the relatively high time consumption of this approach affects the real-time implementation. For some crowd behaviours such as panic dispersing, the abnormality often involves the sudden change of pedestrian's motion patterns such as velocity. Therefore, the drastic change of global motion may indicate the crowd abnormality. In this section, an approach which doesn't consist of complex pattern extraction and classification is introduced to achieve the real-time detection of crowd panic dispersing. The general procedure of this approach is

illustrated in Figure 5-8.



**Figure 5-8. The proposed framework of panic crowd behaviour detection approach**

Once the video data is obtained from the stream, the pre-processing procedure is adopted on the primitive image data. In this research, the noise removal process is implemented. Then the background subtraction will be applied to reduce the computational time consumption of optical flow extraction. As the third step, the optical flow field will be calculated for each frame using the HS optical flow method. The obtained optical flow field will be modelled into a pattern value. In the final step, a trained threshold will be utilized to verify whether the current frame is abnormal. In the following paragraphs, each process step will be further explained in detail.

### 5.4.1 Pre-processing and Parameter Setting

Assuming the length of the raw video is  $F_l$ , the initial several frames of the entire video is selected as the training stage, which is marked as  $F$ . The selection of  $F$  will depend on the actual situation. In this research, the first 20% of the  $F_l$  is selected. The threshold  $T$  will be used to verify whether the current frame is abnormal. In the pre-processing stage,  $T$  will be estimated. Also, the  $T$  is dynamically adjusted as the frame index increases. In order to achieve the noise removal process, the wavelet denoising will be applied on each raw image. In order to balance the time consumption of the optical flow calculation, the rough background subtraction technique is applied. In the feature extraction stage, if the pixel is marked with background, it will not be processed to reduce the computational burden.

### 5.4.2 Feature Extraction and Post Processing

Once the pre-processing is complete, the Horn-Schunk optical flow algorithm is applied on two consecutive frames. For example, if the current frame  $t$  is being evaluated, the  $k$ th  $t + 1$ th frames will be used for the optical flow extraction. The obtained optical flow is noted as  $u_t$ . Once the flow map is obtained, the post-processing will be implemented to optimize the flow distribution. A neighborhood average filtering is implemented on the  $u_t$  to obtain a smoother flow map. The filtered map is denoted as  $v_t$ . The Figure 5-9 illustrates the comparison between the extracted  $u_t$  and the filtered  $v_t$ . The first figure shows the  $u_t$ , and the second figure shows the  $v_t$ . The second figure illustrated the neighborhood average filtered optical flow field. The motions are more fluent and explicit.

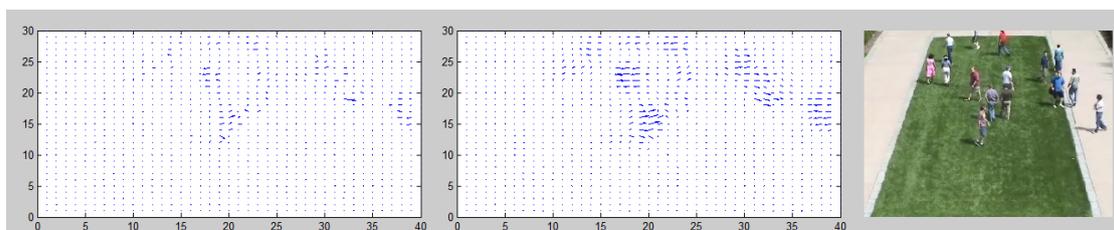


Figure 5-9. A comparison of optical flows before and after the neighborhood average procedure

### 5.4.3 Signature Modelling

Once the pre-processing and feature extracting stages are complete, patterns will be modelled into signature for the detection. When panic dispersing happens, the magnitude of pedestrian's velocity will experience a greatly increasing. Figure 5-10 illustrates the comparison between the optical flow field before and after the crowd panic is triggered. The first figure shows the distribution when pedestrians are walking in a normal state. The second figure shows the flow distribution after the anomaly occurs. When the anomaly happened, pedestrians will run toward to the right instead of walking. The flow map is visually denser than the one in normal state. Since the magnitude of global flow could be simple enough for analysis in the macroscopic perspective, it will be accumulated for each frame as the signature for the panic behaviour detection.

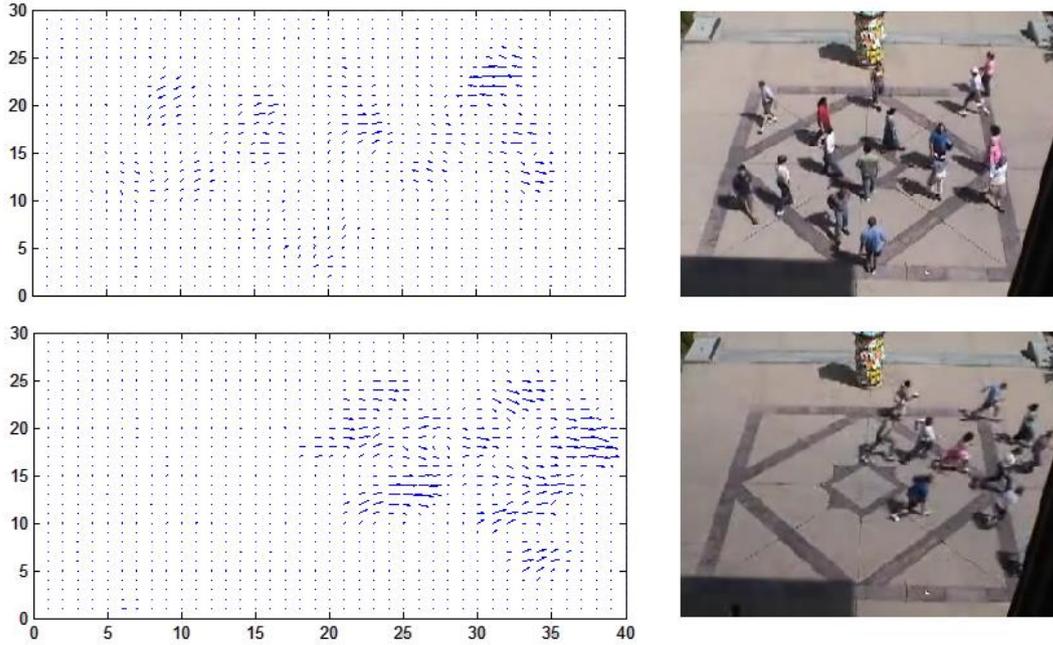


Figure 5-10. Changing of Magnitude in Panic Event

The modelled signature could be expressed as the Equation 5-22. The  $v_k$  is the neighborhood average filtered optical flow field.  $W$  denotes the maximum width of  $v_k$ , and  $H$  denotes the maximum height of  $v_k$ .

$$S = \sum_{w=1}^W \sum_{h=1}^H |v_k^{w,h}| \quad 5-22$$

#### 5.4.4 Model Training

The last step is to determine whether the current frame  $k$  contains panic behaviour, once the signature is modelled from the optical flow. In training stage, the value of threshold  $T$  will be estimated. The estimation of  $T$  could be achieved with two different approaches. In the first approach, the threshold  $T$  is set to a fixed value. The disadvantage is the adaptiveness would be hampered. In the second approach,  $T$  could be dynamically adjusted according to the actual situation. The advantage is the high adaptiveness. However, the challenge of this approach is to devise a mechanism to obtain the optimal  $T$ . In this research, in order to estimate the threshold based on the actual situation, the  $S$  value of first  $F$  frames are accumulated. Then the average value  $A$  of all  $S$  is calculated. The procedure could be expressed as  $A =$

$AVG(S_1, \dots, S_F)$ . Then, the value of threshold  $T$  is two times of  $A$ . Once the process is complete, a reasonable  $T$  could be estimated for the footage. The Figure 5-11 illustrates the difference between the calculated average value  $A$  for the training stage and  $S$  of each frame. In this stage, pedestrians remain in a casual state with low velocity. Therefore, the calculated difference is constant and stable. Thus, the obtained value of  $T$  is 20 in this case.

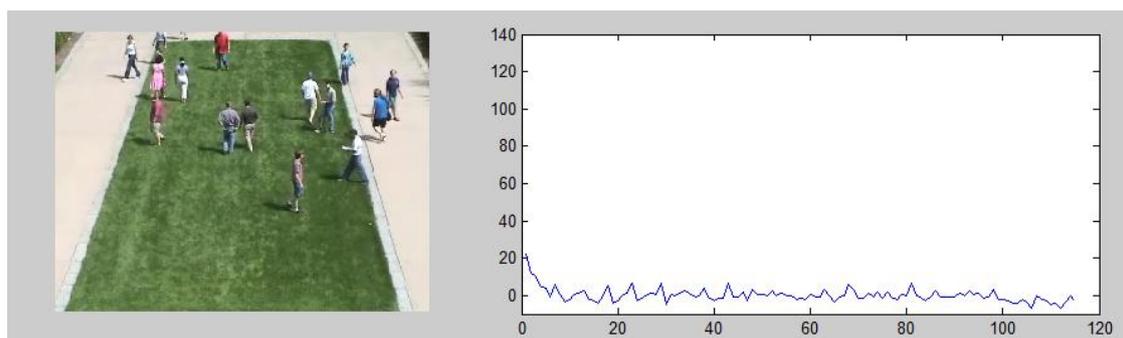


Figure 5-11. Difference values  $s$  between  $S$  and  $A$  value for each frame

### 5.4.5 Anomaly Detection

With the modelled parameters such as  $A$  and  $T$ , frames for testing will be evaluated whether the global magnitude suddenly becomes too large. In the detection phase, if the difference between  $S$  of current  $k$ th frame and the average value  $A$  is larger than the trained threshold  $T$ , this frame will be determined as abnormal. The condition could be noted as  $|S_k - A| > T$ . In Figure 5-12, the trend of  $|S_k - A|$  from the normal stage to the abnormal stage is illustrated. In the example, normal stage starts from the zero frame. After the 450<sup>th</sup> frame, pedestrians begin to disperse in panic state. The figure indicates the magnitude of  $S$  drastically increases when the dispersing begins. The maximum difference could be as large as 120, which is six times larger than the threshold  $T$ . Therefore, the value of  $T$  can be dynamically set larger in order to minimum the false positive detection. The results demonstrated the effectiveness and superior performance from the devised operational pipeline.

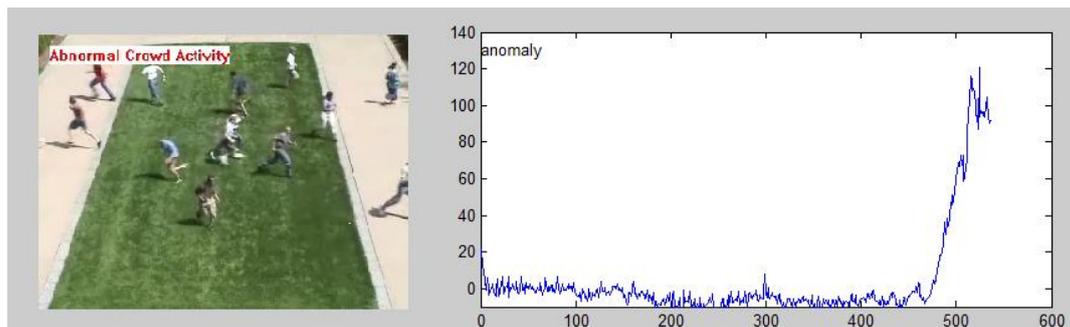


Figure 5-12. Detection results using the proposed framework

The Pseudo Code of this approach is illustrated as below.

---

```

V ← Data                                % Obtain the video data
F = length(V)                            % Obtain the length of video
Ft = 0.2 * F                          % Get the first 20% frames
ST = 0                                % Initialize the Summation
For k = 1: Ft                          % For each frame in the 20% frames
    uk ← OF(V(k))                       % Calculate the optical flow
    vk ← Background_Subtraction(uk)    % Background subtraction

    ST = ST + ∑w=1W ∑h=1H |vkw,h|    % Get the summation
End
A = ST/Ft                              % Calculate the average
For k = Ft + 1: F                      % For the rest frames
    uk ← OF(V(k))
    vk ← Background_Subtraction(uk)

    S = S + ∑w=1W ∑h=1H |vkw,h|
    IF (|Sk - A| > T)                  % If the difference is larger than T
        return Abnormal                % Return abnormal
    End
End

```

---

Listing 5-2. Pseudo Code of the change detection approach

In this chapter, the extraction process of GLCM is introduced and the derived statistical patterns are defined. By analysis their features on the different types of STTs, the most representing patterns are modelled as the descriptor for the classification. Another panic detection approach based on motion magnitude proposed in the early phase of this program is introduced as well.

## **Chapter 6. Complex Crowd Behaviour Synthesis and Simulation**

In this chapter, an innovative model for the crowd behaviour simulations is introduced. This model consists of three fundamental components including long-term path finding, short-term local optimal motion steering and social force-driven interaction handling.

## 6.1 Hybrid Rules for Crowd Synthesis

In this research, various types of crowd behaviours and their extraction techniques are introduced. Actually, these theoretical models and operational principles can be reversed and utilized in crowd simulation. For the sake of visual realism of crowd simulation, 3 aspects need to be taken care of, long term direction (destination), short term motion (contextual and scene awareness) and in-crowd interaction (following, collision avoidance, etc.).

To develop a prototype crowd simulation system, firstly, the classic A-star (A\*) algorithm is adapted as the long-term path finding model in order to achieve the global path planning. Secondly, for the short-term local optimal motion control, a steering algorithm introduced in the research of Reynolds (1987) is exploited to handle the short-term motion decision. Thirdly, for the interaction handling model, an enhanced Social Force Model (Helbing *et al.*, 1995) is renovated to handle the interaction between neighboring agents.

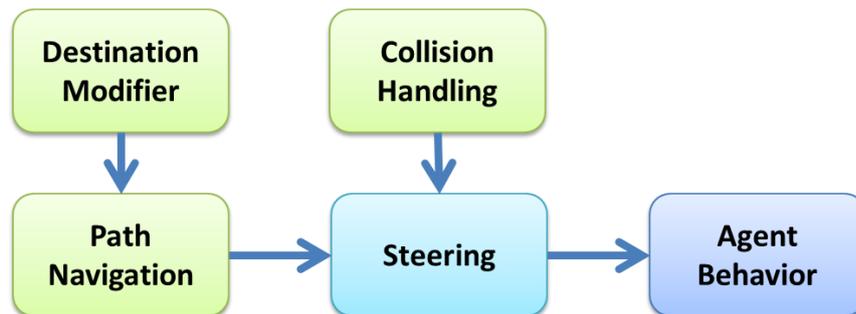


Figure 6-1. The framework of proposed crowd synthesis technique.

As Figure 6-1 illustrates, the path navigation component represents the long-term path finding model based on the A-star algorithm. The destination modifier component serves as the short-term local motion steering model. Take the example of a circling crowd, since they aren't moving toward the destination and their instant velocity and direction is always changing, the short-term steering model adjusts agent's orientation for each frame to keep it on the right path. Furthermore, if an agent is pushed away from its optimal path by interaction force, the short-term steering model will attempt to drag it back. Next, the interaction model serves as the collision handling component

illustrated in Figure 6-1. Its main function is to eliminate the collision between neighboring agents when moving. Usually interaction control models such as social force are utilized as the collision handler. However, due to their defects such as oscillation, an extended Social Force Model is adapted to provide a better visual realism. The above-mentioned components work together to steer the actual motion of agent. Under the influence of the combined components, the devised approach is capable of controlling each agent's behaviour and ultimately lead it to the destination.

### 6.1.1 Baseline Works

In the proposed simulation approach, the A-star path finding algorithm plays a primary role of long-term path navigation. This algorithm is initially introduced by Hart *et al.* (1968). The A-star is an expansion of Dijkstra's algorithm and outperforms its performance on path finding. According to the research of Yang *et al.* (2017), the A-star algorithm is the most efficient path finding method when the road map is static. Furthermore, the A-star algorithm can also play as the heuristic algorithm of solving many other problems. Without implementing any preprocessing, the A-star is capable of directly finding the path within several iterations. In the road network with fixed weights, the A-star algorithm exhibits a better performance. The fundamental idea of A-star algorithm is similar to the Expectation Maximum approach, which is to find the partially probed path with minimum weight and then estimate the distance to destination on every iteration as illustrated in Figure 6-2. The A-star algorithm could be expressed as Equation 6-1.

$$f(n) = g(n) + h(n) \quad 6-1$$

where  $n$  is the index number of current map node.  $g(n)$  is the calculated distance between initial node to node  $n$  in current iteration.  $h(n)$  is the estimated distance between current node  $n$  to destination, the Euclid distance is used as the heuristic function in this research. The scene is a two-dimensional plane and divided into grid of  $m$  by  $n$  blocks or nodes. The height and width of each block is also one unit. In order to calculate the value of  $g(n)$  at node  $n$ , eight neighboring nodes will be measured to get the minimum weight.

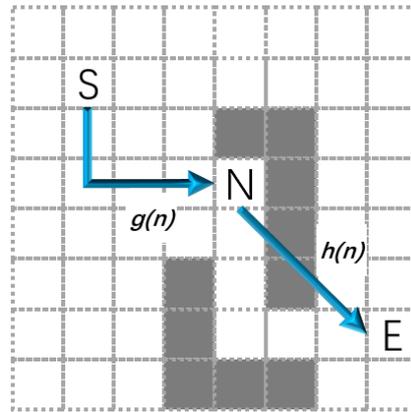


Figure 6-2. The grid of map nodes.

The detailed procedure of A-star algorithm can be expressed as follows. Assume the start node is  $S$ , destination node is  $E$ . For each node  $P$ , the moving consumption from  $S$  to  $P$  could be denoted as  $G_P$ , the distance between  $P$  to  $E$  could be denoted as  $H_P$ . The moving consumption from node  $P$  to node  $N$  could be denoted as  $D_{PN}$ , the processing priority is denoted as  $F_P$ .

- 1) Select the  $S$  and  $E$ , insert  $(S, 0)$  into the open list. Where 0 is the  $F_P$  of  $S$ . The open list is a priority queue, the lower  $F_P$  denotes the higher priority.
- 2) If the open list is empty, the  $E$  is unreachable. Otherwise, pop the  $P$  with lowest  $F_P$ .
- 3) Travel through the neighboring nodes of  $P$ , for each node  $N$ , if  $N$  is in the close list, it will be ignored. Otherwise, it is processed as follows.
  - a) If  $N$  isn't in the open list, let  $G_N = G_P + D_{PN}$ , estimate the distance  $H_N$  from  $N$  to  $E$ , let  $F_N = G_N + H_N$ . Set the parent node of  $N$  as  $P$ , insert  $(N, F_N)$  into open list.
  - b) If  $N$  is in the open list, let  $G'_N = G_P + D_{PN}$ , if  $G'_N < G_N$ , then  $G'_N = G_N$  and recalculate  $F_N$ . Then replace the  $(N, F_N)$  in open list. Set the parent node of  $N$  as  $P$ .
- 4) Put node  $P$  into close list. If  $P$  is  $E$ , the searching is complete. Recursively looking for the parent node of  $P$  until find  $S$ . The path is found. Otherwise, repeat step 2.

The flow-chart of the operation pipeline is illustrated in Figure 6-2.

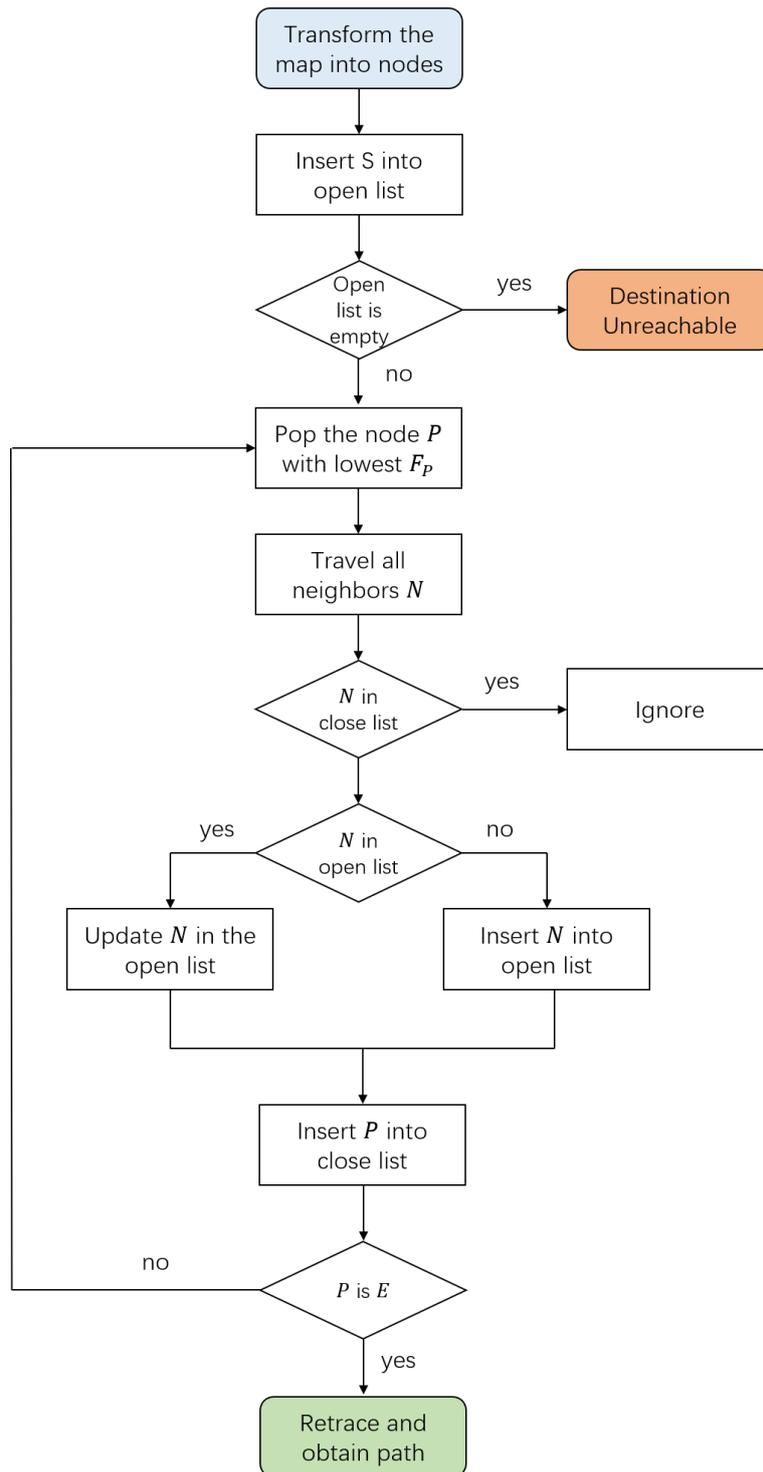


Figure 6-2. Flow Chart of A-star path finding algorithm

### 6.1.2 Personal Space and Relative Velocity

The main function of the collision handling component is to prevent the collision between neighboring agents. To be specific, a repulsive force will be triggered if two agents are getting too close. On the contrast, if the distance becomes larger than the

comfortable personal space, the repulsive force shouldn't take effect any longer. In conventional approaches, the Social Force Model is most frequently exploited. The conventional SFM is introduced in chapter 2.

However, the desired social force  $f_d$  could only handle the crowd behaviour in the environment successfully with simple structure. If the environment in the stage becomes complicated, simulated agents relying on SF without path finding capability would be easily stuck by obstacles. In order to address this issue, the A-star path finding algorithm is adopted to achieve the long-term steering as the Path Navigation Component. And  $f_d$  would serve as the short-term local optimal steering component. For the collision handling model, the two repulsive forces are exploited to prevent the collision between agents and obstacles. Together, these three components ensure the agents from reaching the final destination and crowd behaviours with correct visual realism.

Despite the conventional SFM has been widely utilized in the crowd simulation, several disadvantages expose as further explored. These disadvantages could degrade the visual realism of the crowd, even with illogical behaviours. The first disadvantage is that simulated agent in the conventional SFM is always considered as a rigid body. Despite the repulsive force tries to avoid the collision, it will inevitably occur and the pedestrians should be deformed. In the research of Helbing *et al.* (2002), a physical contact force consists of body force and sliding friction force is imported into the SFM to address this issue. Also, in the conventional SFM, the calculation of repulsive force is based on the relative spatial location of agents. However, this approach isn't always accurate. For example, if a pedestrian is following another, the front pedestrian shouldn't be influenced by the repulsive force if he is unaware of the other pedestrian. Also, agents with different velocity should be psychologically affected by different repulsive forces. In order to address this issue, the repulsive force of the conventional SFM is improved. Two novel concepts include the Personal Space, and the Enhanced Repulsive Force are adapted to enhance the visual realism of the proposed simulation approach.

In the original SFM, the repulsive force between any two agents in the stage will constantly exist. With logarithmic function, if the distance between two agents become large, the repulsive force will become insignificant. This property brings one issue. In the crowd with high-density, for any two agents with large distance, the repulsive force is so small which can't affect the agent's actual motion. However, this process still consumes the computational resource. Overall, it derives large burden to the simulation. In order to address this issue, the Personal Space is introduced to the repulsive force modeling. The influence of Personal Space could be expressed as Equation 6-2.

$$\vec{f}'_{ij} = \begin{cases} \vec{f}_{ij}, & d_{ij} - r_j \leq \rho_i \\ 0, & \text{otherwise} \end{cases} \quad 6-2$$

Where  $\vec{f}'_{ij}$  is the repulsive force in conventional SFM. The constant value  $\rho_i$  represents the Personal Space of agent  $i$ . If the difference between distance and radius is larger than the Personal Space, the repulsive force will be ignored. This process is illustrated in Figure 6-3.

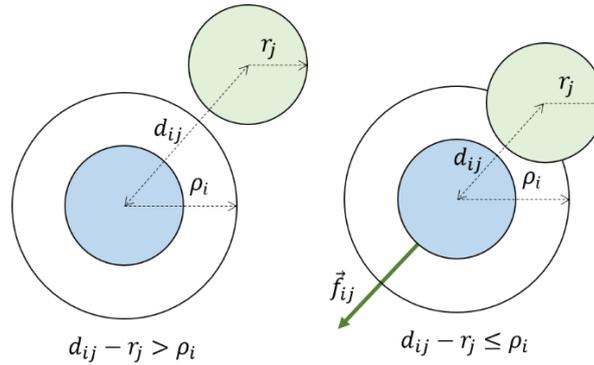


Figure 6-3. The Repulsive Affected by Personal Space

The second concept is the Enhanced Repulsive Force. Because the pedestrian walking in the queue head won't be affected by the repulsive force since he/her doesn't notice others. Therefore, if the pedestrian doesn't notice the one behind him/her, the repulsive force doesn't take effect at all. The more the pedestrian in front notices the pedestrian at back, the higher repulsive force is. This assumption could be expressed as Equation 6-3.

$$\vec{f}_{ij}^{rep} = \begin{cases} \theta(h_{ji}^n) \vec{f}'_{ij}, & d_{ij} - r_j > \rho_i \\ (1 + \theta(h_{ji}^n)) \vec{f}'_{ij}, & \text{otherwise} \end{cases} \quad 6-3$$

Where  $h_{ji}^n$  is defined as  $h_{ji}^n = (\vec{v}_j - \vec{v}_i)\vec{n}_{ij}$ , which is the Relative Velocity. Under its influence, if two agents are moving with the same direction, the force magnitude will be smaller. However, if they move in the opposite direction, the force would increase. The value of function  $\theta(z) = z$ , if  $z > 0$ . Otherwise,  $\theta(z) = 0$ . Under this situation, if the function  $\theta(z)$  detected two agents in opposite directions are about to collide, the repulsive force will take effect. But if these two agents already passed by each other, the repulsive force will be ignored.

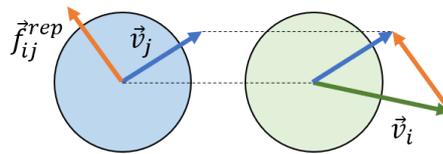


Figure 6-4. The Repulsive Force affected by Relative Velocity

Combined with the A-star path finding algorithm as the long-term steering component, the final social force would be determined as Equation 6-4.

$$m_i \frac{d\vec{v}_i}{dt} = f_{A^*} + \sum_{j \neq i} \vec{f}_{ij}^{rep} \quad 6-4$$

Where the  $f_{A^*}$  denotes long-term desired force if the agent is on the right track of the optimal path obtained by the A-star algorithm. If not, the  $f_{A^*}$  denotes the short-term desired force to drag the agent back to the correct path. By mapping the script with this proposed behaviour model to each agent, visually realistic crowd footages could be synthesized. The simulation experiments will be conducted in the chapter of experiments.

### 6.1.3 Enforced Group Social Force Model

In previous sections, a crowd modelling approach consisting of long-term path finding, short-term steering and interaction handling models is devised. This approach has addressed the global and local behaviours of each simulated agent. However, the common social relationship is ignored. In real-life, pedestrians familiar with each other will attempt to stay close. Despite the local interaction is fully composed of the

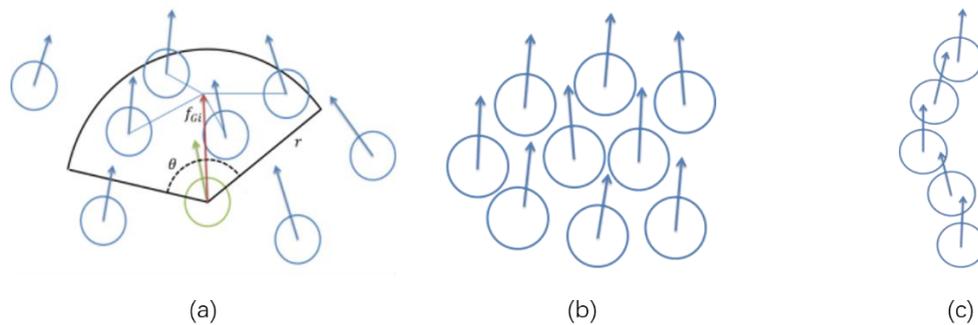
repulsive force in the previous research, the attracting force of SFM is omitted. Therefore, the research of this section will further expand the proposed model, in order to simulate the interactions between agents with common social relations.

#### a) Group Attraction Force

Inspired by the Cohesion rule of Boids Model introduced in previous chapter, pedestrians knowing each other will be psychologically kept closer by the force when moving toward the same destination. Therefore, a novel Group Attraction Force is introduced to simulate this phenomenon. The Group Attraction Force affecting agent  $i$  is pointing to the center of spatial position of all agents within agent  $i$ 's perception field, and marked as  $f_{Gi}$ . The  $f_{Gi}$  could be defined as Equation 6-5.

$$f_{Gi} = A_i \log_e \frac{DIS(AVG(C_{a \in G}), C_i)}{B_i} n_{Gi} \quad 6-5$$

Where the  $C_a$  represents the spatial position of any agent with the same destination within the perception field of agent  $i$ . Notation  $G$  is a collection of all agents with same destination index. The equation  $DIS(AVG(C_{a \in G}), C_i)$  indicates the distance between agent  $i$  and the spatial position center of all agents in  $G$ . The notation  $n_{Gi}$  indicates the direction of Group Attraction Force, which points from  $C_i$  to  $AVG(C_{a \in G})$ . The  $A_i$  and  $B_i$  adjust the magnitude of  $f_{Gi}$ , their values depend on the actual situation. According to this equation, it could be notified that if agent  $i$  is far from the average center of  $G$ , the magnitude of  $f_{Gi}$  will become larger, and vice versa. The calculation of  $f_{Gi}$  is illustrated in Figure 6-5(a).



**Figure 6-5. The Field Perception and the impact on result with different parameters, (a) The Perception Field of an agent  $i$ . (b) Grouping result when Field Perception is Broad. (c) Grouping result when Field Perception is narrow**

### b) Perception Field

The Perception Field is another important concept to affect the group behaviour. It assumes that the perception of each agent can't cover everyone on the stage. The Figure 6-5(a) illustrates the Perception Field. The Perception Field is a Fan-shape area with the center at the spatial position of agent. The notation  $r$  is the perception radius which determines how far the agent could sense. And the notation  $\theta$  is the perception angle which determines how wide the agent could see. Other agents covered within this perception field will be sensed, on the contrary, other agents will be ignored. In the figure, only four agents will be noticed by the agent  $i$ . The import of perception field could significantly affect the shape of simulated crowd. When the value of  $\theta$  is set to a large value, the formation of crowd could become wider, as illustrated in Figure 6-5(b). On the contrast, if the value of  $\theta$  is set to a smaller value, the crowd will be narrow as Figure 6-5(c). In the simulation, if the formation of the crowd become too wide or narrow, the visual realism of simulated video would be hampered. Therefore, an appropriate setting of perception field is crucial for the simulation.

### c) Enforced Crowd Simulation Model

With the Group Attraction Force and Perception Field, the proposed agent behaviour model is updated as Equation 6-6. The  $gsf_i$  notes the final affected force applied to agent  $i$ . This force is determined by three factors, which are long-term steering force  $f_{A^*}$ , interaction force  $\vec{f}_{ij}^{rep}$  and the group attraction force  $f_{Gi}$ . With the updated model, the proposed approach is expected to have a better performance while simulating the agents with various destinations.

$$gsf_i = f_{A^*} + \sum_{j \neq i} \vec{f}_{ij}^{rep} + f_{Gi} \quad 6-6$$

## 6.2 Prediction using the Enhanced Social Force Model

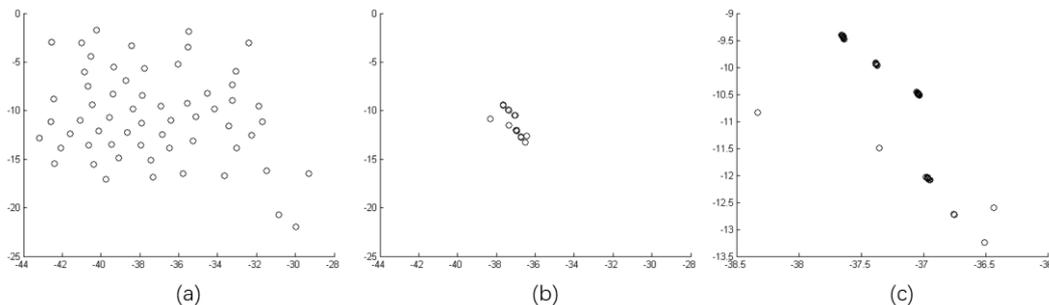
According to the definition of the proposed Group Attraction Force, forces derived from agents with same destination will point to a common center. This center has the

potential to be utilized as an identifier to backtrack the destination of agents. In this section, a method to predict the destination of agents from two different groups is introduced based on the Group Attraction Force.

The trajectory prediction usually involves with the probability models such as Hidden Markov Model. These prediction models rely on the spatial patterns. However, the extraction of spatial patterns for each agent in high crowd density is often inaccurate. Especially when agents are clustered in a crowd, and affected by the intensive interaction forces. Since the group attraction forces  $f_{Gi}$  of agents with same destinations point to the average center. According to the definition of Cohesion rule, it could be assumed that average centers of agents within the similar view perspective would cluster as one. If these clusters are spatially separated, they could be segmented to verify which group the agents belongs to.

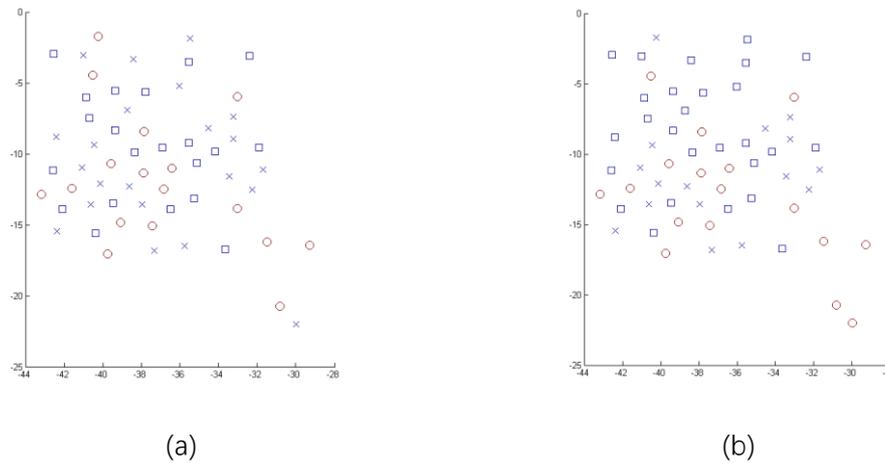
### 6.2.1 Assumption Validation

Experiments are conduct to validate the proposed assumption. The spatial position of agents and average centers are collected from the crowd simulation, and illustrated in Figure 6-6. In the video, 30 agents from 3 different groups are generated. As illustrated in Figure 6-6(a), the spatial positions of agents are randomly distributed. It is difficult to determine which group the agent belongs to. The Figure 6-6(b) illustrates the distribution of average centers obtained from the simulation. The average centers are clustered under the influence of the group attraction force. After magnified the scale of Figure 6-6(b), the distribution of average center is illustrated in Figure 6-6(c). The average centers are clustered according to the group index of agents.



**Figure 6-6. The comparison between agents' distribution and the extracted grouping centers. (a) Distribution of agents. (b) Distribution of grouping centers. (c) Center distribution in magnified scale**

The Figure 6-7(a) illustrates the actual ground truth of the agents' distribution. 30 agents from three different groups are labelled with different shapes, which are round, square and cross respectively. The Figure 6-7(b) illustrates the prediction result using KNN clustering algorithm. The result implies the accuracy is relatively high. Therefore, the calculated average center is utilized as a novel pattern for the crowd's common destination prediction, namely Grouping Center and will be marked as  $AC_i$ , where  $i$  is the index of the agent.



**Figure 6-7. Comparison between ground truth and segmentation result. (a) Ground Truth. (b) Segmentation Result using proposed pattern**

## 6.2.2 Predict results on the simulated crowd

In order to further validate the prediction performance, the proposed approach is implemented on ten simulated crowd footages. For each footage, three groups of agents are generated, and each group consists of 20 agents. Since the initial spatial positions of the agents are randomly distributed, the simulated videos vary for each iteration. The simulated video is composed with two different stages. For the first stage, agents are randomly mixed. And for the second stage, controlled by the proposed behavioural model, agents with same destination will eventually clustered. In the second stage, the clustered agents could be easily separated with the spatial information. In order to assess

the prediction capability of proposed approach, the prediction uses video data of the first stage. The Grouping Center  $AC_i$  will be calculated with the data collected from a certain frame. The prediction accuracy for each iteration is recorded, and once all ten simulations are implemented, an average accuracy will be calculated. In the experiment, the extracted Grouping Centers are segmented with the conventional K-mean clustering.

The Table 6-1 illustrates the prediction accuracy of the ten simulations. Eight of ten experiments achieved the 90% plus accuracy. The fourth, sixth and seventh runs of the experiment reach 100%. However, the results of eighth and ninth runs obtain low accuracy. The reason is part of the grouping centers formed an additional cluster. Overall, the average accuracy of these ten simulations reached 88.83%, which proves accessibility of the proposed approach.

<b>Runs</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>Accuracy</b>	91.67%	93.33%	93.33%	100%	93.33%
<b>Runs</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
<b>Accuracy</b>	100%	100%	60%	63.33%	93.33%
<b>Average</b>	88.83%				

**Table 6-1. segmentation accuracy on simulated videos**

### 6.2.3 Structure of the behaviour prediction approach

Under the premise that the extracted  $AC_i$  of all agents are detected, the Grouping Centers could be easily acquired. However, data used in last paragraph is collected from the simulation. The spatial position and group attraction force could be precisely exported from the simulation tool. But in the real-life footage, the Grouping Centers  $AC_i$  needs to be modelled. According to the Equation 6-5, the Grouping Center  $AC_i$  could be estimated with the group attraction force and agent's spatial position. The agent's spatial position could be obtained with the pedestrian detection approach. However, the group attraction force is also unknown. Equation 6-7 could be exploited to estimate the group attraction force  $f_{Gi}$ . Once  $gsf_i$ ,  $f_{A*}$  and  $\vec{f}_{ij}^{rep}$  are obtained, the

$f_{Gi}$  could be calculated. Therefore, the main objective is to find an efficient way to obtain  $gsf_i$ ,  $f_{A^*}$  and  $\vec{f}_{ij}^{rep}$ .

The Figure 6-8 illustrates the proposed procedure for crowd prediction. Agents are detected in the first stage. Detected agents will be utilized to estimate the repulsive force, actual velocity and desired force. Next, the group attraction force is calculated, and the grouping centers are obtained and clustered. Through the backtracking process, the destination of agent could be predicted.

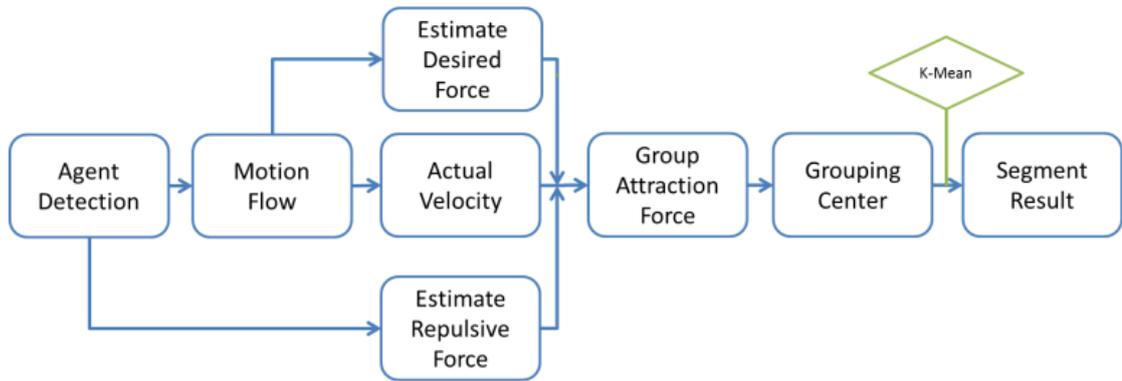


Figure 6-8. The framework of the proposed crowd prediction approach

Assume the actual affected force, long-term desired force and repulsive force are correctly calculated and estimated, the group attraction force  $f_{Gi}$  could be obtained using the following formulation derived from Equation 6-7.

$$f_{Gi} = gsf_i - f_{A^*} - \sum_{j \neq i} \vec{f}_{ij}^{rep} \quad 6-7$$

In order to get the agent's destination, the point of interest approach is adapted to find out the possible destination in the scene. For example, if the motion flow density is significantly higher than rest of the scene for a long term, there might exist a destination such as exit. Therefore, this region could be considered as a point of interest. Once all the points of interest are obtained, the track-let of agent will be extracted from the flow map and the belonging of the point of interest will be evaluated to determine the direction of desired force.

The actual velocity is the actual motion state of agent. The global flow-based features will be utilized for the calculation. The conventional optical flow extraction

algorithm such as HS and LK will be applied to obtain the optical flow map. Next, each agent will be tracked with a pedestrian detection approach. Then, the flow map and the distribution of agents will be matched, and each agent's actual affected force could be estimated. The actual affected force is calculated from the particle mapped to the agent and eight neighboring particles, as shown in Equation 6-8.

$$gsf_i = k_1 f_{x,y} + k_2 (f_{x-1,y-1} + f_{x-1,y+1} + f_{x+1,y-1} + f_{x+1,y+1} + f_{x-1,y} + f_{x,y-1} + f_{x+1,y} + f_{x,y+1}) \quad 6-8$$

Where  $f_{x,y}$  indicates the actual affected force to agent  $i$  at the co-ordinate  $x$  and  $y$ . The  $k$  is the weight factor, because the force at the center position has a stronger impact on determining the final force, the value of  $k_1$  is set to 1. When calculating the neighboring values, the value of  $k_2$  is set to 0.4. The estimation of repulsive force  $\vec{f}_{ij}^{rep}$  is using the enhanced SFM which is introduced in section 6.1.2.

Thus, the methodologies for the calculation and estimation of these three forces are introduced. According to the equation, the Group Attraction Force  $f_{Gi}$  could be obtained. Because  $f_{Gi}$  is essential a vector, along with the agent's spatial position  $(x_i, y_i)$ , the Grouping Center  $C_{x,y,i}$  will be finally acquired. And they will be fed to the classifiers to segment. According to the clustering results, the corresponding agents will be backtracked and the destination will be mapped to the detected point of interest.

In this chapter, a crowd simulation approach consists with long-term and short-term agent behaviour control models is proposed. The A-star path finding algorithm exploited by the long-term control model is firstly discussed. Then, the concepts of personal space and relative velocity are adapted for the short-term interaction handling model. The group attraction force model for the group behaviour handling is introduced as well. Inspired by the proposed group attraction force model, a behaviour prediction approach is also proposed and the preliminary experiment is conducted to prove its accessibility.

## **Chapter 7. Experiments and Evaluation**

In this chapter, experiments are conducted with the proposed approaches in previous chapters to validate their performances. The contents are distributed as follows: Section 7.1 introduces the benchmarking video datasets for crowd behaviour analysis based on varied crowd density; Section 7.2 exhibits the effectiveness of the STT extraction method and the performance of the devised behaviour recognition approach based on GLCM against other benchmarking techniques. Section 7.3 demonstrates the experimental results on the crowd panic dispersing detection model. Section 7.4 concludes the crowd simulation results using the devised crowd simulation system. Section 7.5 elaborates the crowd behaviour prediction results.

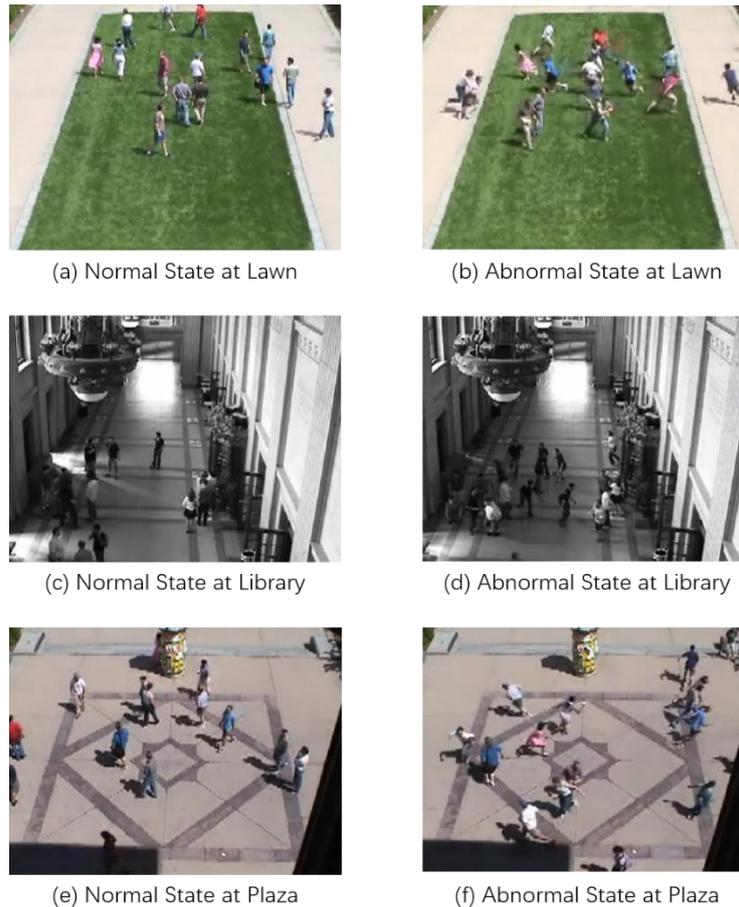
## 7.1 Datasets for Crowd Behaviour Analysis

In this research, commonly used benchmarking crowd video datasets for behaviour analysis are deployed. The crowd density is a crucial factor of selecting the analyzing techniques. Most techniques having supreme performance on the crowd behaviour recognition in normal-density but a fast-declining performance when applied in challenging real-world environment such as those taken from high density outdoor scenes. Hence, the datasets in the experiment are divided into two, medium and high-density crowds.

### 7.1.1 Datasets with Medium Crowd Density

#### a) The UMN dataset

The UMN dataset is collected from the campus of the University of Minnesota, and all footages are performed based on predefined scripts. Videos from this dataset contain three scenes, which include lawn, plaza and library. Three video clips are performed by dozens of pedestrians for each scene. Figure 7-1 illustrates snapshots of footages in UMN. Figure 7-1(a) and Figure 7-1(b) illustrate snapshots of normal and abnormal states at a lawn ground. In Figure 7-1(a), fifteen pedestrians are walking casually on the lawn. In Figure 7-1(b), pedestrians are dispersing in a panic state. Figure 7-1(c) and Figure 7-1(d) illustrate snapshots at a library. When the anomaly happens, pedestrians are escaping along all directions. In Figure 7-1(e) and Figure 7-1(f), snapshots of videos taken from a plaza scene are illustrated. In Figure 7-1(e), pedestrians are walking in a normal state. In Figure 7-1(f), when the anomaly happens, pedestrians escape in left and right directions.



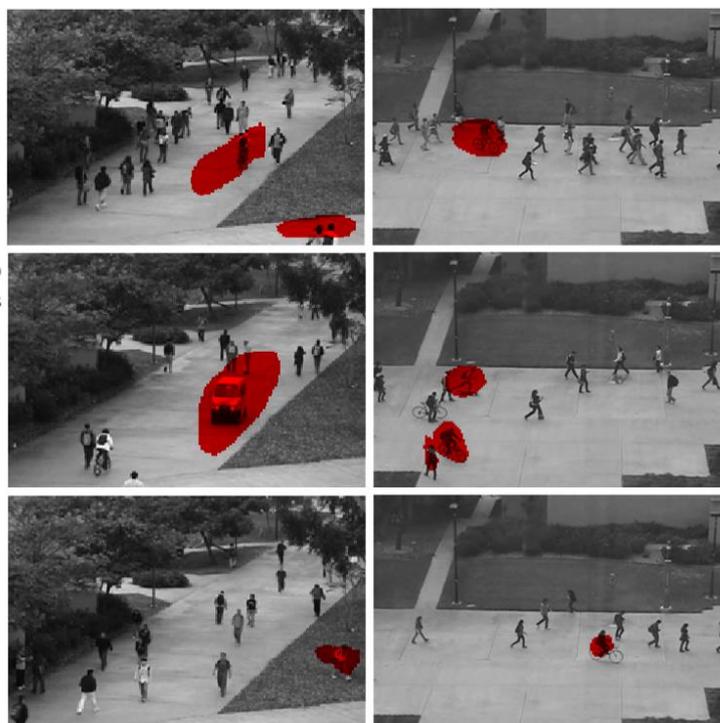
**Figure 7-1. Normal and abnormal snapshots from UMN dataset**

The UMN data set is widely used in lab-based experiments (the criticism is it is based on simplified/idealized setting against the often much more challenging actual field trial scenarios). Cui *et al.* (2011) devised an Interaction Energy Potentials (IEP) model and tested it on the UMN. Ma *et al.* (2014) proposed an algorithm via online learning, and implemented it on UMN to compare with the performances from using the SFM, Optical Flow model, Streak line Potentials model, and Sparse reconstruction model. Raghavendra *et al.* (2011) exploited UMN to test the proposed Optimizing Interaction Force algorithm. Venkatesh and Zhu (2012) test the proposed locality model on UMN and compare the performance to Chaotic Invariants, Social Force, Optical Flow, and Sparse methods.

#### b) The UCSD Dataset

The UCSD Anomaly Detection Dataset is acquired from a stationary camera

mounted at an elevation, overlooking pedestrian walkways. The crowd density in the walkways ranges from sparse to very crowd. Abnormal events are caused by either: 1) the circulation of non-pedestrian entities in the walkways or 2) anomalous pedestrian motion patterns. The snapshots of the UCSD dataset are illustrated in Figure 7-2.



**Figure 7-2. Snapshots from UCSD dataset and the labelled anomalies**

#### c) PETS 2009 Dataset

The PETS2009 Dataset contains multi-sensor sequences of different crowd activities. It is composed with five subsets: 1) calibration data; 2) training data; 3) person count and density estimation data; 4) pedestrian tracking data; 5) flow analysis and event recognition data. Each subset contains several sequences, and each sequence contains different view perspectives (from 4 up to 8).

#### d) Forensic Dataset

The forensic video dataset collected from real-life criminal events are also considered as an important data source for crowd analysis. This dataset contains 50 videos from the record of real forensic criminal cases. The Figure 7-3(a) illustrates the snapshot of a video contains the violent behaviour near a construction set. In this video,

construction workers are trying to violently fight against the police, and the police fires shots to alarm the crowd. In Figure 7-3(b), a snapshot from surveillance camera records a severe fight bursts between two groups of users. The Figure 7-3(c) illustrates a fight at the basketball playground at US. All of these videos are obtained from real-life forensic evidence which is collected from the CCTV cameras installed in the public area. Therefore, this video dataset could be utilized to evaluate the possibility of the practical implementation for the proposed crowd analysis approach. However, a significant pattern of this dataset is that the density of the crowd is still not high enough for the analysis of extremely dense crowd.



Figure 7-3. Snapshots from forensic video dataset

### 7.1.2 Datasets of High Crowd Density

The density of crowd in UMN, UCSD, PET2009 and forensic dataset is relatively low. In the research of crowd analysis, the crowd with high density is widely required. The approach exhibits good performance on low density crowd may generate unsatisfying detection result on high density crowd. Therefore, video datasets with high crowd density is also introduced.

In the research of Cheney (2004), crowd videos collected from football matches are collected for a dataset. The dataset contains both violent and non-violent crowd behaviours. This video dataset contains 125 video footages with normal cheering

behaviour in football matches and another 125 video footages with violent fighting behaviours. Each video has the length of five to ten seconds. The following Figure 7-4 illustrates snapshots of the dataset. In Figure 7-4(a), a snapshot of audience cheering for their football team is illustrated. Note that despite most of the crowd exhibit drastic motion, the crowd behaviour is non-violent. For Figure 7-4(b), football fans from different groups are fighting with each other. In this case, crowd might have the same motion magnitude with Figure 7-4(a), but its behaviour contains violence. The challenge is how to recognize the difference between these videos. On the other hand, these videos consist of the crowd with high density, which could be exploited for the analysis of extremely crowded situation.



**Figure 7-4. Snapshots from dataset with extreme high density**

In the experiments of the crowd behaviour analysis approaches, these video datasets will be exploited. Note that the simulated crowd videos are also been utilized.

## 7.2 Classification Results using GLCM Signature

In this section, experiments are conducted to assess the performance of devised signature modelled from the GLCM matrix for the recognition of two crowd abnormal behaviours including panic dispersing and congestion. The proposed procedure of experiment is illustrated in Figure 7-5. As mentioned in the previous section, the six-directional Gabor Filtering is applied on the target STT extracted from STV. After the background is filtered, the STT is divided into patches of  $m$  by  $n$  pixels. For each patch, the calculated signatures are sent to the classifiers, either to train or determine

what behaviour they belong to. Other features such as TAMURA proposed by Ranjan *et al.* (2016) will also be used for the crowd behaviour detection, and the experimental results between the proposed signature and TAMURA will be compared.

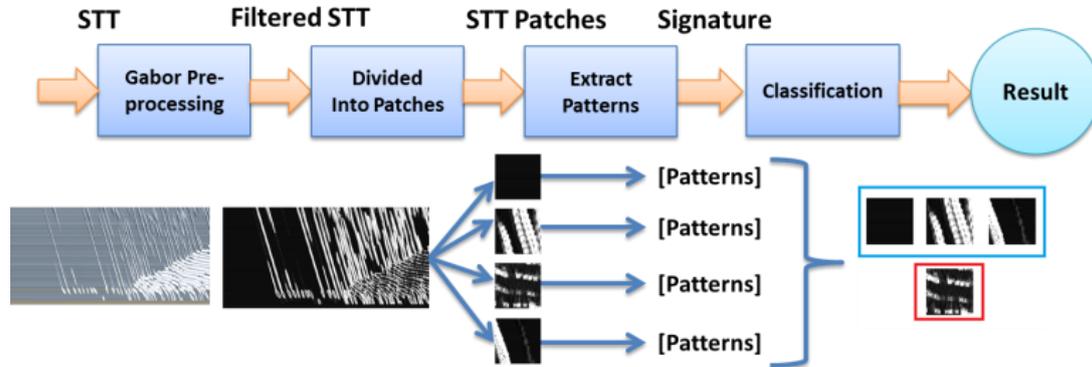


Figure 7-5. Structure of proposed classification approach

### 7.2.1 Training Process

In the experiment, various types of classifiers are exploited for the performance evaluation of the proposed signature and TAMURA. Five different classifiers are used for the analysis including the K Nearest Neighbor (KNN), Naïve Bayes, Discriminant Analysis Classifier (DAC), Random Forest and Support Vector Machine. In the training process, STT patches are manually labelled as four different types including Empty, Normal, Congested and Panic. For the Empty patch, no motion trajectory exists except the static background. For the Normal patch, pedestrians are walking in a normal state, and the motion trajectory's slope is small. For the Congested patch, the velocity of pedestrian is low and the trajectory's slope is even smaller. Also, the motion strips are more condensed. For the Panic patch, the velocity of pedestrian is large and the slope is also large. Once labelled, the patches will be used for the training of the classifiers.

### 7.2.2 Recognition Result

Once the training is complete, STT for classification will also be divided into patches, and corresponding signature will be modelled. The parameters for classifiers are set according to the following regulations. The  $m$  and  $n$  of the patch are set to 50 and 50 in value. For the KNN classifier, the  $k = 4$  includes Empty, Normal, Panic and

Congestion respectively. The default setting will be implemented to other classifiers such as Naïve Bays, DAC and SVM in MATLAB. As illustrated in Figure 7-6, the classification result of KNN on the proposed signature is shown. Grids in blue color indicate the patch border. The cross mark in white color indicates the patch with Empty state. The cross mark in green color indicates the patch with normal state. The cross mark in amber color indicates the patch with congested state. Visually, patches representing different behaviours will show different patterns. When the crowd is congested, pedestrians' motion would be very slow. Under this circumstance, pedestrian's spatial shifting will take longer time. This will result the stripes or trajectories to have a slope with smaller value in congested state. On the other hand, if the crowd isn't at the congested state, the extracted STT will exhibit condensed parallel stripes with small slope value. In summary, patches contain congested crowd behaviours will show trajectories with smaller slope than the one contains normal state. Furthermore, Contrast, Entropy and Variance of patches with congested behaviours usually have smaller value.

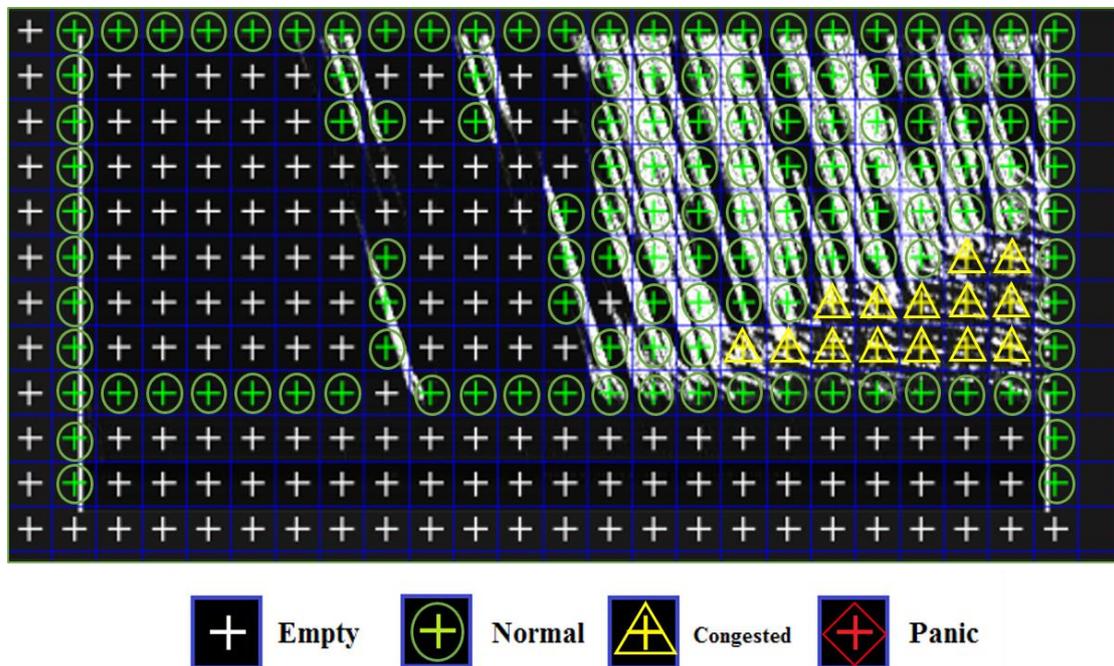


Figure 7-6. Detection result using GLCM signature and KNN

As the comparison of performances on different feature, the TAMURA is also been exploited with the same procedure of behaviour classification. As illustrated in Figure

7-7, some patches with zero motion trajectory is labelled with green cross. Also, some patches with trajectory in normal density and slope are labelled with amber cross. This result indicates that the proposed signature has a higher accuracy on the congestion recognition.

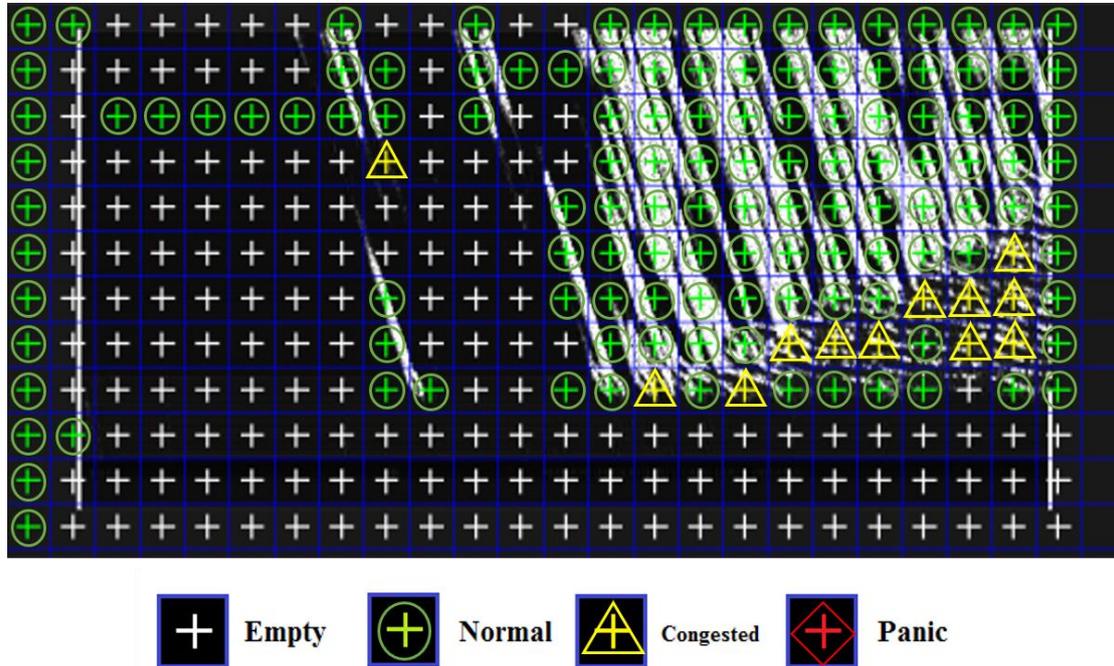


Figure 7-7. Detection result using TAMURA signature and KNN

The performance of recognizing the panic dispersing crowd behaviour is also evaluated. The classification results are illustrated for both features in Figure 7-8. The Figure 7-8(a) shows the result using proposed signature, and Figure 7-8(b) shows the result using TAMURA. Both of the experiments have applied KNN classifier on the features. It could be observed that the performance of TAMURA is slightly better than the proposed signature for this certain video.

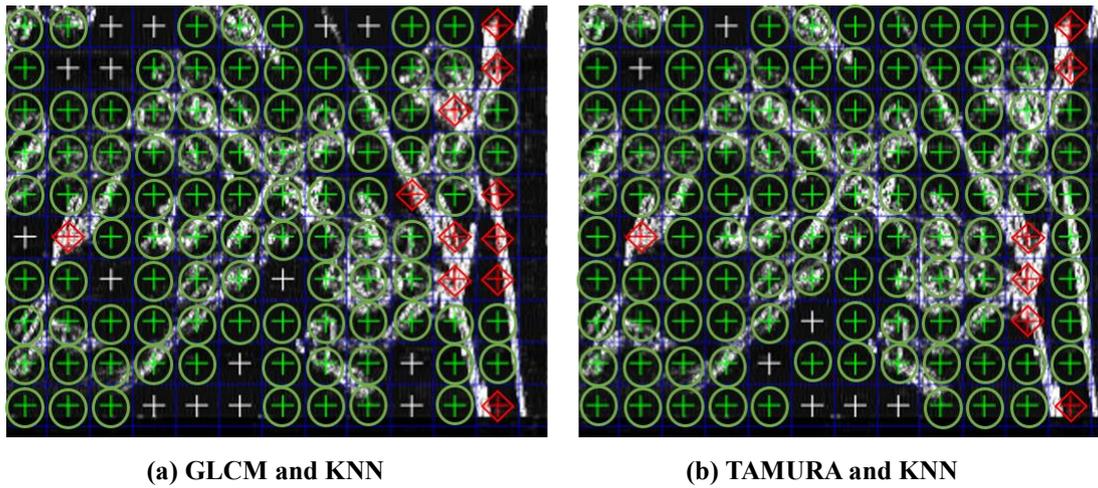


Figure 7-8. Comparison of detection results on panic dispersing

In order to further evaluate the performance of the proposed signature on different classifiers, the different combination of signatures and classifiers are implemented on all 11 videos in UMN set, UCSD set, Forensic set and 4 simulated video footages built with simulation tool. The average accuracy on all videos is calculated for each combination as the final score. For each patch, if the determined behavioural type equals to the manually labelled ground truth, the detection is considered successful. Thus, this patch  $C_{i,j}$  is marked as either 1 or 0. The overall accuracy  $A$  for each video is obtained with Equation 7-1. And the accuracy for each feature/classifier combination is listed in Table 7-1.

$$A = \frac{\sum_{i,j=0}^N C_{i,j}}{i * j} \quad 7-1$$

Feature	Classifier	Simulated	UMN	UCSD	Forensic
GLCM	KNN	75.14%	72.38%	82.31%	85.71%
TAMURA	KNN	78.86%	65.24%	83.79%	81.03%
GLCM	SVM	85.03%	70.12%	86.72%	88.37%
TAMURA	SVM	61.31%	70.12%	85.29%	82.10%
GLCM	Naïve Bayes	85.39%	55.24%	79.69%	81.10%
TAMURA	Naïve Bayes	76.34%	68.27%	77.13%	79.93%

GLCM	DAC	76.67%	66.37%	75.38%	76.26%
TAMURA	DAC	82.65%	70.89%	78.49%	74.82%
GLCM	RandomForest	83.38%	65.54%	82.34%	75.29%
TAMURA	RandomForest	81.37%	70.00%	79.19%	79.36%

Table 7-1 Accuracy of multiple signatures and classifiers combination

### 7.3 Panic Dispersing Detection Results

In the experiments, all 11 video clips from the UMN dataset is analyzed using the proposed change detection approach on panic dispersing. The detection results are illustrated as Figure 7-9. The line in blue denotes the detection result, and the line in red denotes the manually labelled ground truth. When the value is zero, the current frame is normal. Otherwise, the frame is panic. Overall, all panic dispersing behaviours are successfully detected, besides some minor exceptions. For the second video, the detected panic appears several frames before the labelled ground truth, it is because labelling of the ground truth is not accurate. Also, the detection quickly becomes normal, because once pedestrians left the camera, the value of  $S$  becomes small, the current frame will be considered as normal. But for the ground truth, current state is still abnormal. In order to address this problem, one possible solution is using a dynamic threshold which could be updated according to the actual changing of  $S$ . Another issue is the false positive detection such as the 10th detection. This problem is because the inappropriate value setting of thresholds. The threshold could be set larger to avoid this issue.

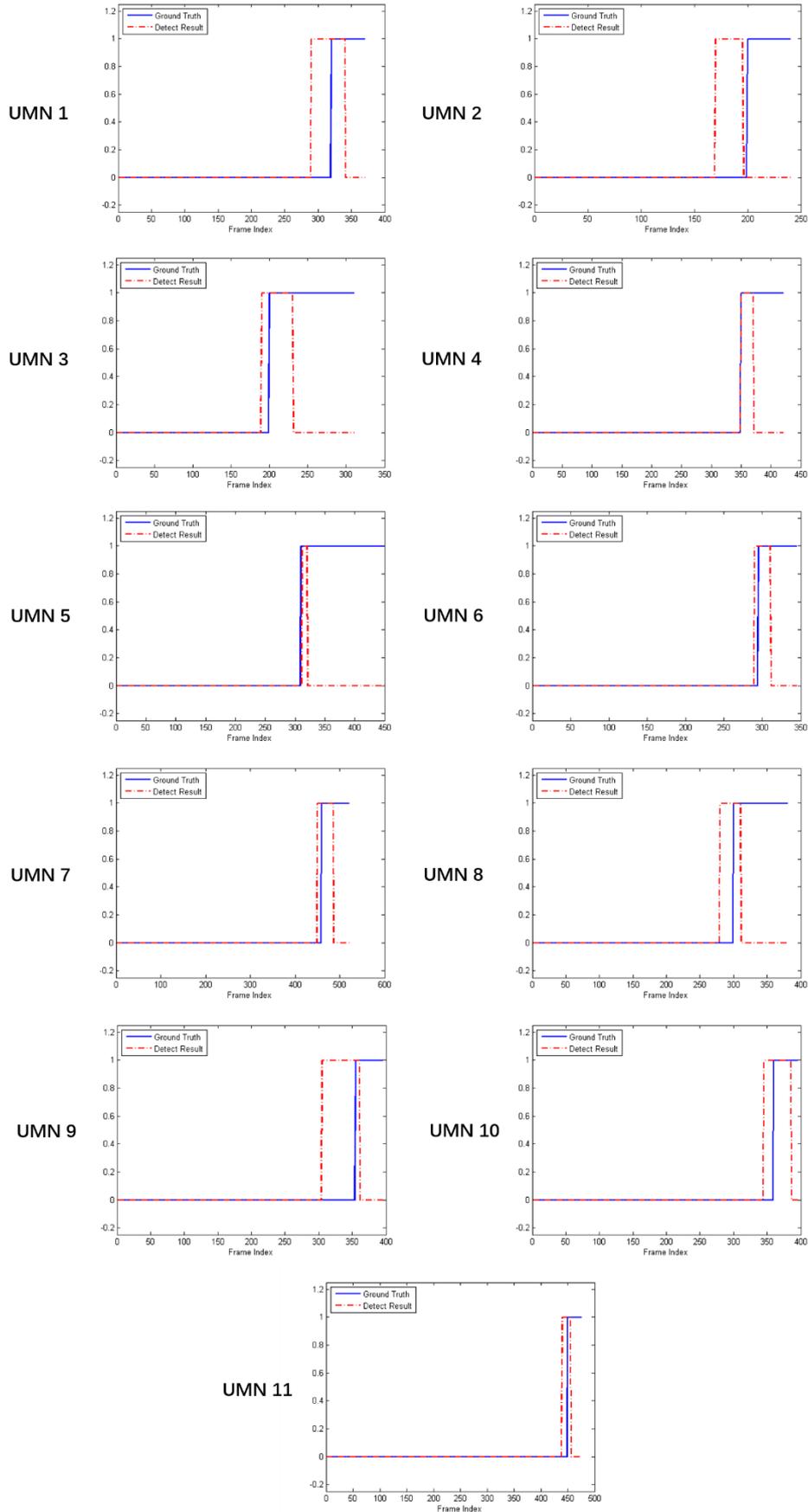


Figure 7-9. Detection result of the proposed change detection approach

## 7.4 Game Engine based Simulation

In this section, six different types of crowd behaviours with an additional hybrid behaviour in the proposed taxonomy will be simulated.

### 7.4.1 Simulation Tool

For all simulations, the three-dimensional game engine Unity will be exploited for the crowd behaviour simulation. The engine Unity 3D is a multi-platform game development tool for developers to create interaction contents such as 3D video games, visualization of constructions and real-time 3D animations. Similar to tools such as Director, Blender game engine, Virtools and Torque Game Builder, Unity provides a GUI environment for the developing. Its IDE runs at Windows and Max OS, and is capable of publishing games to platforms such as Windows, Mac, Wii, iPhone, WebGL, Windows phone and Android. Overall, it is a powerful graphical tool.

### 7.4.2 General Installation for All Simulations

For all simulated videos, basic installation is set up for the simulation tool. The installation consists of three processes. 1) Agent generators. This model assigns the frequency, spatial distribution and initial parameters of generated agents. The proposed behavioural model is code with C# as a script file, then mapped to an object of agent as a prefab. When the simulation begins, the generator loads the prefab of agent, duplicates it, and generates the crowd. Multiple generators could exist simultaneously to spawn agents with different behavioural patterns. 2) Destinations. The destination doesn't have to be an instantiated object, it might be fabricated into the agent's behaviour model. The spatial position of destinations determines the long-term and short-term desired force, which is a very important component. 3) Environment. The environment plays the role of vessel to contain the generated agents. In simulations, the three processes will be adjusted to generate crowd behaviours including Lane, Crossing, Circling, Dispersing and fountainhead. A collection of simulated videos is illustrated in Figure 7-10.

### 7.4.3 Installations of Different Crowd Simulations

- Bottleneck Behaviour

To simulate the entrance of a building, two obstacles are placed with a narrow space between them. And the destination is set to the right side of obstacles. The generator replicates agents at the left side of the footage, and its spatial position follows the gaussian distribution. During the simulation, the typical pattern “faster is slower” for bottleneck crowd behaviour is clearly exhibited. By setting the generating rate of agent generator to a larger value, agents will be jammed at the entrance. Under the influence of repulsive force, agents with higher velocity would take more time to pass the entrance.

- Lane Behaviour

Two walls at the top and bottom sides of the footage are modelled as the passage. Two agent generators are set in the simulation process. The first generator spawn agents from the left side of scene, and the generated agents attempt to move forward to the destination located at the right side. On the contrary, the second generator is located at the right side of the footage, and generated agents is moving to the destination at the left side. Different agents are rendered with different textures. In the simulated crowd, the Lane Effect shows explicitly. Under the influence of repulsive force, agents from the same group automatically line a queue while moving.

- Crossing Behaviour

No environment installation is required in this simulation. One generator locates at the left side and another one locates at the top side of the footage. Agents from different generators are rendered with different textures as well. In the simulation, crowd successfully exhibits patterns such as Bypassing and Avoidance for the imminent collision.

- Panic Dispersing Behaviour

In this simulation, a generator produces agent from the left side of the footage and moving to the right. The danger source rendered with different texture is modelled at the center of scene. At the normal state, agents are affected by the repulsive force, and

the danger source is treated as an obstacle. After switching to the abnormal state, agents attempt to escape from the danger source, and the collision avoidance effect is greatly hampered since the repulsive force no longer takes the dominant influence.

- Circling Behaviour

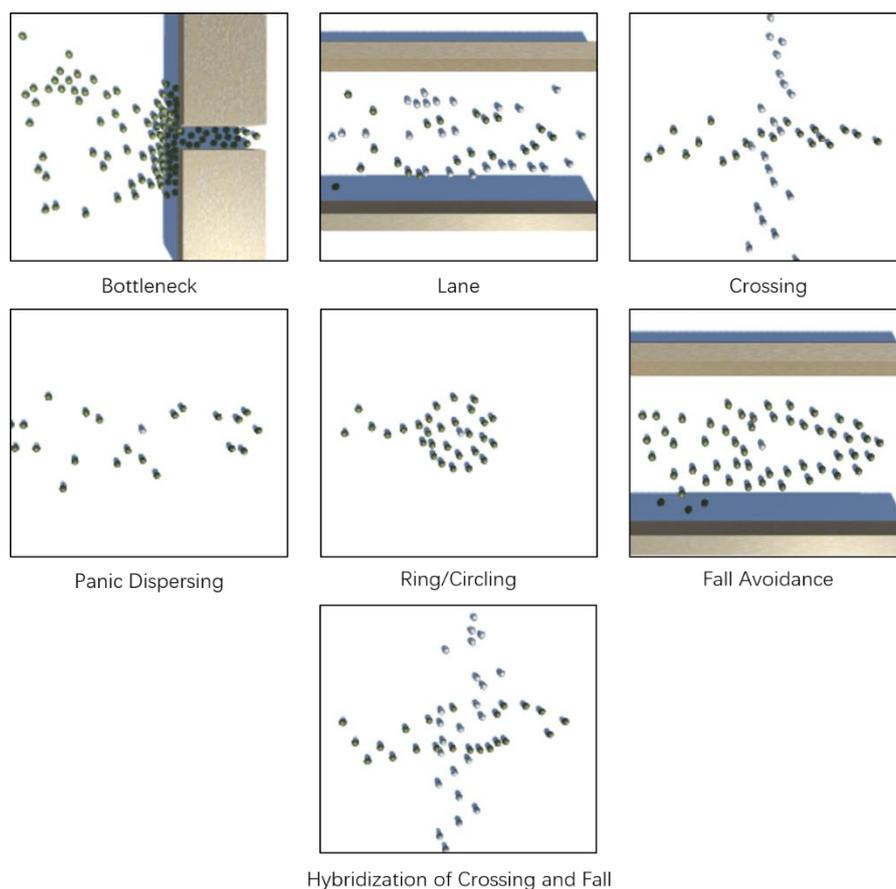
In this simulation, the direction of desired force  $f_d$  is usually pointed to the destination. However, the direction of  $f_d$  will always be set perpendicular to the current velocity. The spatial position of destination is set at the center of footage. The simulation result shows that all agents are circling around the destination. On the other hand, agents maintain a good balance of distance under the influence of repulsive force.

- Fall Avoidance Behaviour

To simulate this behaviour, one agent is placed at the center of the footage as the fallen pedestrian. Its behaviour model is removed so it won't be affected by the social force. The simulation result implies that the crowd flow exhibits the separate and remerge phenomena under the impact of desired and repulsive force.

- Hybrid Behaviour

For the experiment of hybrid behaviours, Fall Avoidance and Crossing are combined. The simulation result implies that both patterns from Fall Avoidance and Crossing are recreated.



**Figure 7-10. Snapshots of simulation results using proposed approach**

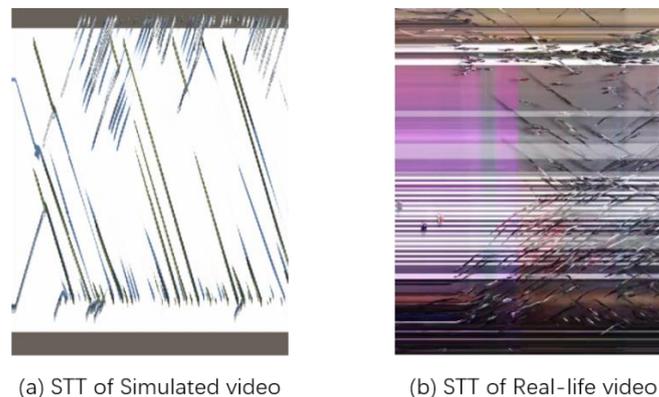
#### 7.4.4 Evaluation of Simulation Performance

Despite behaviours are successfully simulated, the realism of the simulated crowd behaviour is still unknown. The most crucial factor is the visual realism because the primary purpose of crowd simulation is to be as much similar to the real-life as possible. For all the researches at recent decades, there are no benchmarking standards to evaluate the visual realism of the simulated crowd. The approaches for the evaluation have a great variety. For example, Stuart *et al.* (2015) uses surveys on a large pool to evaluate the general acceptance of the visual realism. Firstly, a website with the devised survey system is built up and opened to the public. The concept of Two-Alternate Force Choice (2AFC) is adopted to the system. The system has collected a large set of simulated video data with both proposed and other algorithms. For each time of choice, two manually selected videos with similar scenarios are shown to viewers to decide which one is more realistic. In the experiment, the result is a statistic by a large pool of

viewers. The advantage of this evaluation approach is the accuracy could be significantly high if viewer pool is large enough. However, the first disadvantage is that each cycle of evaluation would consume large amount of time. Another disadvantage is that the judgement of viewer is always based on the subjective decision. Despite the visual realism is subjective, however the evaluations of same videos would derive different results.

Despite the evaluating pattern of the simulated crowd behaviours is still under exploration, some patterns for the behaviour in smaller scale between the crowd have been proposed. For example, while applying the repulsive force using the original SFM, the simulated crowd will exhibit clear oscillation effect. Therefore, this unrealistic phenomenon could be exploited to measure the performance of simulated repulsive force.

Another possible pattern could be exploited to assess the performance of the simulation algorithm. Patterns extracted from the crowd videos such as trajectories and textures are mainly utilized to classify the different behaviours and abnormalities. In this research, the STT could be exploited for the evaluation. As illustrated in Figure 7-11, the extracted STTs are illustrated for comparison. The Figure 7-11(a) shows a STT from the simulated crossing behaviour, and Figure 7-11(b) is a STTs extracted from the real-life video obtained from the cross road at Tokyo. Note these STTs exhibits the similar visual patterns.



**Figure 7-11. STTs comparison between simulated and real-life scene**

As illustrated in Table 7-2, patches representing different behaviours are manually

selected. For example, the first patch of Table 7-2 is selected from a texture from a real-life video with no pedestrian exists. The fourth to sixth patches are selected from the simulated video which contains situations include no-agent, normal walking and crossing. In the experiment, modelled patterns of GLCM will be extracted from the selected STT patches, and the results are listed in the Table 7-2. By comparing the pattern values between patches with same behaviours of both simulated and real-life video, it could be observed that the changing trend of pattern values roughly matches. For example, the entropy value of patch with no pedestrians is less than the one with pedestrians for both real-life and simulated video. Therefore, these patterns could be exploited for the evaluation of realism as a supporting factor.

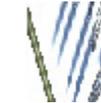
	Empty	Same	Opposite	Empty	Same	Opposite
Patch						
Contrast	0.0310	2.6966	3.2560	0	1.1033	0.6612
Dissimilarity	0.0220	0.9919	1.0889	0	0.4347	0.2938
Homogeneity	0.6623	0.3243	0.3110	0.6724	0.5174	0.5585
Similarity	0.6629	0.3614	0.3512	0.6724	0.5348	0.5680
ASMt	0.6093	0.0359	0.0391	0.6724	0.2740	0.2694
Energy	0.6335	0.1519	0.1562	0.6724	0.4170	0.4114
Entropy	0.1291	2.1923	2.2168	0	1.1915	1.1451
Varriance	0.0356	1.9970	2.5652	0	1.0881	1.6975
Correlation	0.0641	0.2112	0.2351	0	0.3314	0.5335

Table 7-2 STTs pattern value comparison between real-life and simulated videos

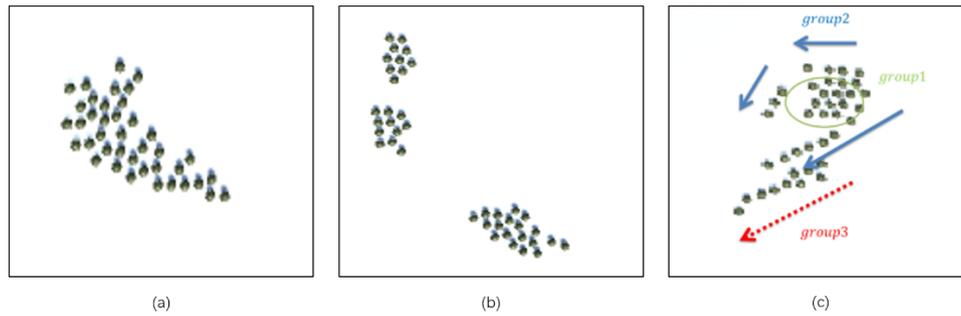
### 7.4.5 Simulation of Crowd with Grouping Behaviour

In this section, the crowd with grouping behaviours will be simulated with the proposed A-star and Enhanced SFM approach. Under the influence of group attraction force, the initially mixed agents would eventually group into clusters. The general

installation remains the same. The terrain is a plane without any obstacles. Note that the plane needs to be wide enough so that agents won't exceed the border. The number of agents is fixed, and will be generated at the initial stage. To achieve the required behaviour, generated agents are expected to exist a longer period. Therefore, agents will move along a circular route.

Next, two experiments with different scenarios will be implemented to assess the performance of the grouping behavioural model. For the first experiment, the main purpose is to verify the influence of Group Attraction Force. In the experiment, three groups of agents are randomly distributed as one single cluster. Each of group consists of 10 agents. Agents are expected to be only influenced by the group attraction force from other agents from the same group. On the other hand, the repulsive force will be generated between any agents. Also, the velocity of agents from different group is set as an identical value to ensure the influence is only from the repulsive and group attraction force. The Figure 7-12(a) illustrates the initial state of the agents after several frames. Note that all agents are clustered together at the initial stage.

After 30 seconds, the formation of the simulated crowd will change significantly as illustrated in Figure 7-12(b). Behavioural patterns are observed under the impact of group attraction force. The first pattern is the clustering of agents by group indices. As illustrated the Figure 7-12(b), agents automatically clustered into three different groups. It could be verified that each cluster consists of agents from the same group. Under the influence of group attraction force, despite the velocity is set to identical value, the agents fall behind will be dragged to the average center. The distance among agents from the same group will gradually become smaller. Eventually, the repulsive force and group attraction force will derive a dynamic balance between agents. This pattern could be an evidence to prove the group attraction force achieves a good performance on crowd simulation.



**Figure 7-12. Simulated Crowd and exhibited motion patterns. (a) Initial State. (b) Grouping result after 30 seconds. (c) The separation and convergence while routing obstacles**

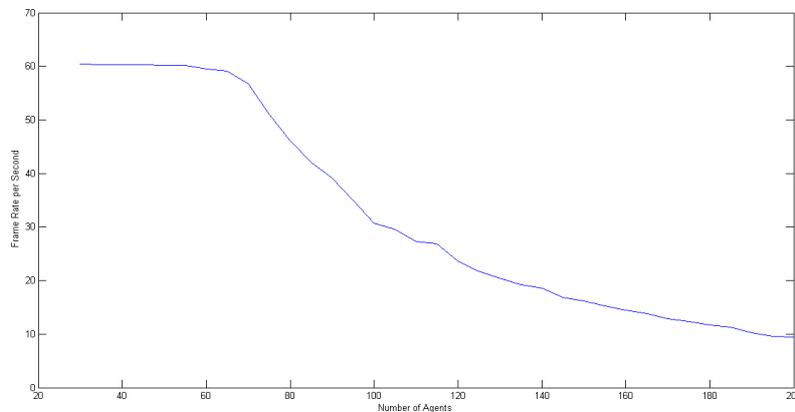
Another behaviour pattern is the automatic maintenance of the cluster formation. The group influenced by the group attraction force would show wide or narrow formation. If the formation becomes too wide or narrow, the visual realism of the simulation might be hampered. Therefore, the parameters of perception field such as  $\theta$  and  $r$  must be carefully set. As illustrated in Figure 7-12(b), the clustered groups exhibit a natural appearance.

For the second experiment, the velocities of agents from different groups are set to different values. The main purpose of this experiment is to assess the performance of the formation distortion and recovery when agents' motion is affected by the obstacles in the stage. The group attraction force should not affect the avoidance behaviour while agent is attempting to bypass the obstacles or other agents. And agents should merge into a cluster again under the influence of group attraction force after bypassing the obstacle. As illustrated in Figure 7-12(c), since the velocity of group 2 is higher, agents separates while bypassing. Then agents from group 2 merges again once passed group 1.

#### 7.4.6 Computational Efficiency

In order to further evaluate the simulation approach, the time consumption is also investigated. In most of cases, the researches of crowd simulation usually support the modelling of agent number in the range from dozens to hundreds. In extreme circumstance, millions of agents will be modelled simultaneously. Therefore, if the computational efficiency is low, the framerate will be seriously affected. Because the

repulsive force needs to be calculated between all agent pairs, assuming the population of agents is  $n$ , the total calculating times for all repulsive forces will be  $n \times (n - 1)$  for each frame. Despite the involving of the perception field and personal space could avoid some computational burdens, the general complexity for each frame is still  $O(n^2)$ . In the experiments, multiple simulations are implemented with the increasing number of agents from 30 to 200, and the average framerate for each simulation is recorded. The experiment results are shown in Figure 7-13. Using the proposed modelling approach, if the number of agents is lower than 70, the simulation could remain the framerate around 60, which is the desired value. When the number is larger than 70, a major drop of the framerate will happen. And a merely 10 framerate remains when the number of agents reaches 200. The experiment result indicates the proposed approach shows a good efficiency while simulating the crowd in mid-high density on the Unity platform. As for the simulation of crowd with high density, the algorithm should be further optimized.



**Figure 7-13. The relation between Frame rate Per-Second and Number of Agents**

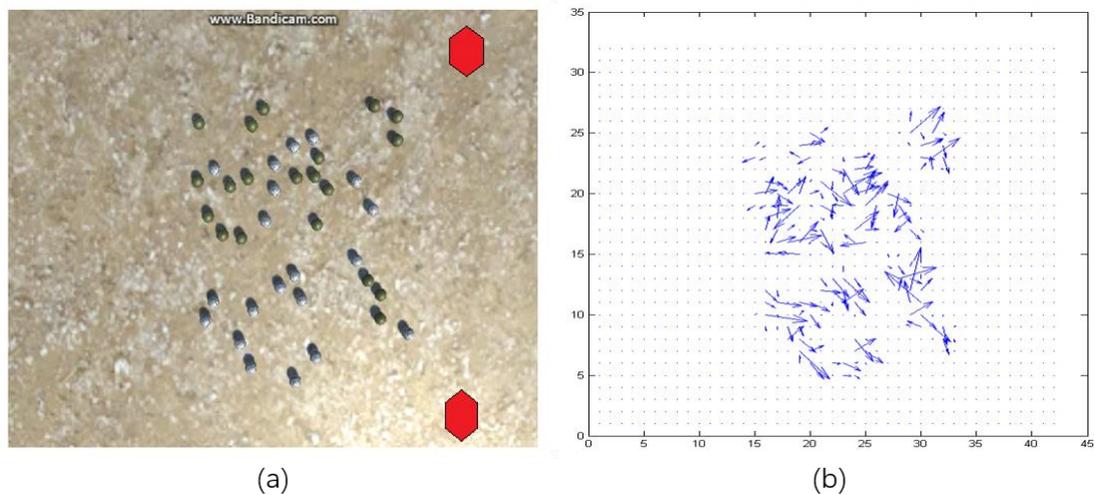
## 7.5 Crowd Prediction Result

In this section, the result of crowd prediction with the grouping center is introduced. According to the proposed approach, the prediction procedure includes the motion flow extraction, pedestrian detection and social force estimation. In section 7.5.1, the

simulated crowd for prediction is introduced. Section 7.5.2 shows the pedestrian detection and motion mapping result. Section 7.5.3 exhibits the procedure of social force estimation and calculation based on the motion map. In section 7.5.4, the prediction accuracy is compared between different classifiers at different frame indices.

### 7.5.1 Crowd Simulation for Prediction

The simulated crowd consists of 40 agents. In order to achieve the higher agent detection accuracy, the agent is a sphere with rigid body mapped with the behavioural model. The crowd contains agents from two groups, each group covers twenty agents. Agents from different groups are rendered with different textures. As shown in Figure 7-14(a), the first group is rendered in green, and the second group is rendered in pale. Agents will be generated simultaneously at the initial stage. Two destinations are marked at the right-top and right-bottom on the footage with the shape of hexagon. The first group is targeting to the destination at the right-top, and the second group is targeting to the destination at the right-bottom.



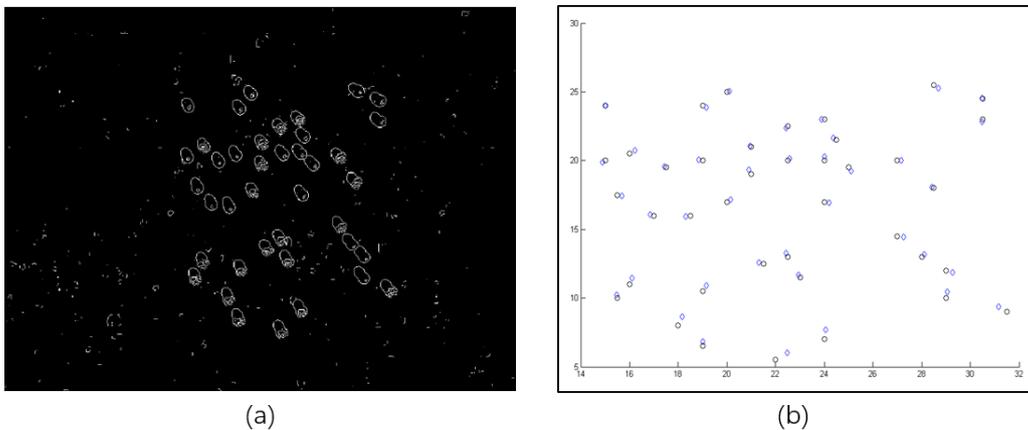
**Figure 7-14. (a) Snapshot of the simulated crowd (b) The extracted optical flow**

Figure 7-14(a) illustrates the initial stage of the crowd. The figure is taken from the 20<sup>th</sup> frame from the simulated video. The simulated agents are at randomly mixed state.

### 7.5.2 Pedestrian Detection and Motion Mapping

The first step of the prediction procedure is to calculate the optical flow field and detect the spatial position of agents. For the flow-field extraction, the 20<sup>th</sup> and 22<sup>nd</sup> frames of the simulated video are utilized. The reason of choosing the frames from the initial stage is the randomly distribution of crowd. After the first several frames, under the influence of grouping force, two groups will be separated. Next in Figure 7-15(b), the HS optical flow is extracted from the image in Figure 7-15(a).

The second stage is to extract the spatial positions of all agents. If the crowd is not in an extremely high density, the regular pedestrian detection algorithm would satisfy the requirement. In this experiment, the Hough circle detector are exploited to fulfill this task. 1) In the first step of Hough circle detection algorithm, the edges of current item in the scene are extracted using the conventional edge detection approach. In this case, the standard Sobel operator is utilized. The detected edges are illustrated as Figure 7-15(a). 2) Assuming the radius of sphere is determined at range from 9 to 12 pixels, a circle function is adapted on all pixels to detect the shape of circle. The parameters of Hough function are set as follows. The radius step is set to 1, angle step is set to 0.1, minimum circle radius is 9, maximum circle radius is 12 and threshold is set to 0.51. The Figure 7-15(b) illustrates the comparison between the ground truth and detected agents with Hough circle detector. The manually labeled agents are marked with the black circle, and the detected agents are marked with the blue diamond shape.

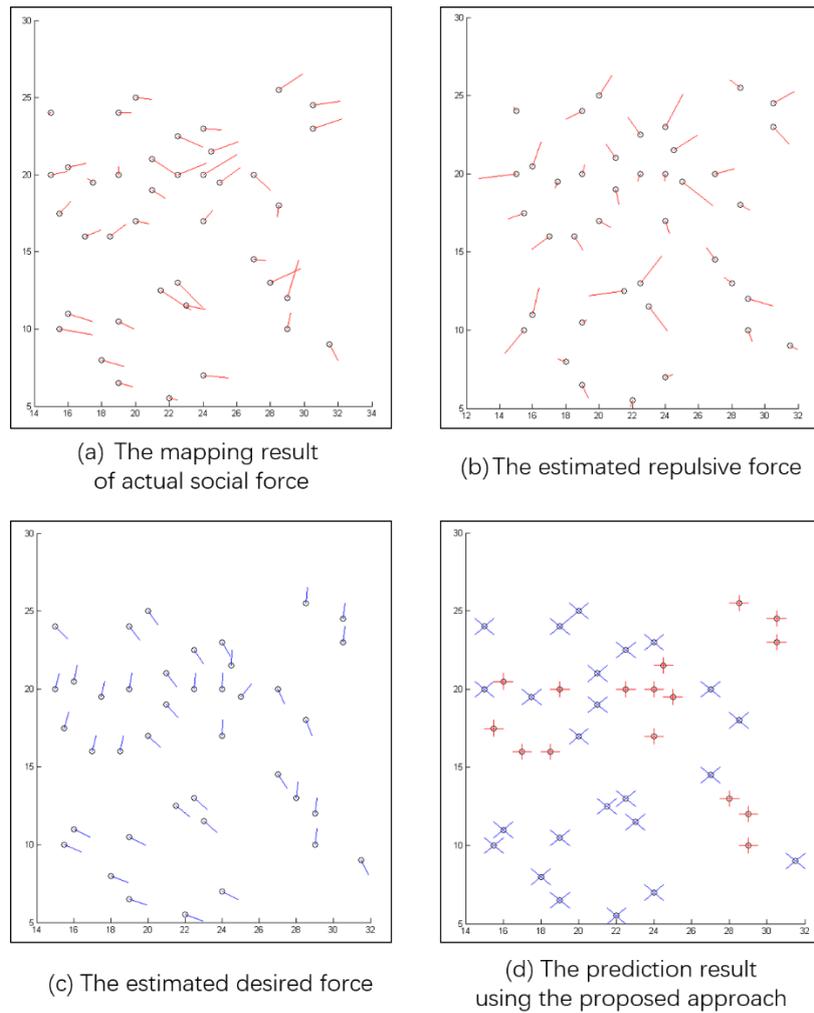


**Figure 7-15. (a) The detected edges. (b) A comparison between location between ground truth and detected agents**

Once the agent's spatial position and the optical flow field are obtained, each agent will be matched to the flow map, in order to assign the actual force affected to the agent. According to the proposed approach, the eight neighboring flow vectors of the agent are accumulated with the center vector, to form the final actual force. The value of weight factor  $k$  for neighboring flow is set as 0.4, and set to 1 for central flow. The mapped force for each agent is illustrated as Figure 7-16(a). Based on the comparison between the simulated video and mapped result, the instant motion of current frame is generally matched with a few deviations.

### 7.5.3 Social Force Estimation

The interaction force between agents will be estimated according to the extracted spatial position of agents. The Figure 7-16(b) illustrates the obtained repulsive force between agents according to the proposed SFM model. The Personal Space  $\rho$  is set to 5 pixels. The Agent Radius is set to 0.5 pixels. The  $A_i$  and  $B_i$  in Equation 2-13 is set to 2 and 0.5 respectively. Note that the  $B_i$  is set to 1 for a more explicit expression.



**Figure 7-16. The estimation procedure and prediction result**

Next, the long-term desired force is the estimated. The point of interest algorithm successfully learnt two destinations from another simulated video. The destination is predicted by tracking the trajectory of agent. But note that the prediction usually didn't match the final result, because the final result is based on the classification of grouping center but not the point of interest. The Figure 7-16(c) illustrates the prediction result of the agent's destination.

After the extraction of actual affected force, the estimation of interaction force and long-term desired force are complete, the Equation 6-7 could be used to calculate the Group Attraction Force  $f_{Gi}$ . Then, the Grouping Center  $C_{x,y,i}$  for each agent will be calculated according to Equation 6-5. In the experiment, the grouping centers are clustered with KNN, and the segmentation result is shown as Figure 7-16(d). Agents

from the first cluster is labelled with the red cross, and agents from the second cluster is labelled with the blue cross. The accuracy is 87.5% for this case.

### 7.5.4 Evaluation of Prediction Result

In this section, the accuracy of proposed approach is compared to various approaches. The prediction accuracy is compared every 5 frames. The result is illustrated in Figure 7-17. In the first experiment, the SVM is applied on the proposed patterns. The training data is generated from the early stage of the video footage. Without manually labelling, all training data are labelled with the result of conventional K-mean cluster algorithm. In Figure 7-17, the blue solid line indicates the accuracy using SVM on the proposed grouping center pattern. It could be observed that the average accuracy from the 5th frame to 110th frame is above 90%. The brown dash line indicates the accuracy using K-mean clustering algorithm on the proposed grouping center pattern. Its segment performance is slightly lower than the previous approach, but still above 80%. For comparison, SVM is applied on conventional spatial position. The accuracy is labelled with solid red line. The K-mean clustering algorithm is applied on spatial position and the accuracy is labelled with green dash line. For the initial several frames, the proposed approach obtained significantly higher accuracy than others. In the late stage, since the agents from different groups are already spatially separated, all approaches achieved high accuracy. Generally, the proposed grouping center exhibits an outstanding performance than the others on the prediction of randomly distributed crowd.

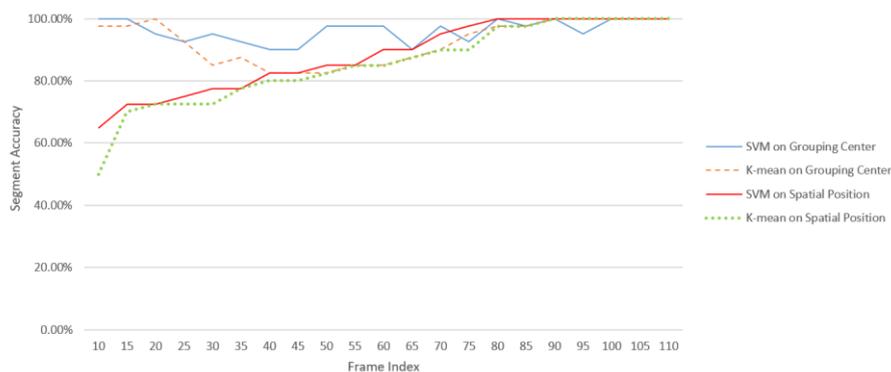


Figure 7-17. A performance comparison between proposed pattern and others

## Chapter 8. Conclusions and Future Work

### 8.1. Contributions to Knowledge

This research focuses on addressing key problems of crowd behaviour analysis and abnormality detection. The discussions encompass the taxonomy of crowd behaviours, behaviour recognition models based on the Spatio-Temporal Texture and GLCM, and crowd behaviour prediction based on the enhanced social force model. These works have delivered to the objectives set at the start of the project and contributed to the domain knowledge on the following aspects.

#### 8.1.1. Explicit Crowd Behaviour Definition and Taxonomy Principles

In this contribution, the definition of crowd behaviour is discussed. An innovative taxonomy is also introduced based on crowd behaviour taxonomies proposed by Somayeh and Robert (2008), Hamidreza and Javad (2016) and Solmaz *et al.* (2012). According to the unique pattern of each behaviour, seven basic behavioural types are defined including Bottleneck, Fountainhead, Ring/Circling, Panic Dispersing, Congestion, Crossing, Avoidance and Lane. For example, the Congestion type consists of patterns such as faster is slower and arching clogging. Crowds with all seven types are simulated with the proposed synthesis approach in the experiment. The experiment results indicate the simulated crowd exhibits the desired behavioural patterns.

#### 8.1.2. Effective Spatio-Temporal Texture Extraction Approach

In this contribution, an effective spatio-temporal texture extraction approach with the enhanced Gabor filter background subtraction is proposed. As the initial step of conventional behaviour recognition procedure, the purposed effective STT extraction is to obtain the texture with the most motion information of pedestrians. The video data is firstly modelled into a spatio-temporal volume, and a collection of STTs are evenly

sampled. According to the unique texture pattern of STT, a six-orientated Gabor filter is devised to subtract its background. After the motion information in STT is obtained, the information entropy is calculated. The highest entropy value indicates the STT with the most motion information, and will be selected as the target STT for the training and recognition processes. In the experiment, the proposed approach is implemented on 7 videos of UMN dataset. The result indicates all extracted STTs contains the most motions of pedestrians. By comparing to the approach without applying the Gabor background subtraction, the result of proposed approach outperforms other approaches on all videos. The main advantage of this approach is low time consumption and relatively high texture quality if the sampling density is set properly. The problem is the decision of sampling density. To address this issue, the STV could be divided into several blocks along time axis before the STT sampling.

### **8.1.3 Novel Crowd Behaviour Recognition Model and Pipeline**

In this contribution, two contributions on crowd behaviour recognition have been made. Firstly, a novel signature modelled with GLCM extracted from the STT is devised to detect the congestion and panic dispersing behaviours of the crowd. Secondly, a change detection approach based on modelled optical flow information is proposed to achieve the crowd panic dispersing with lower time consumption.

According to the proposed procedure of abnormal behaviour analysis, signature will be modelled with the extracted STT. Then the modelled signature will be used to train the classifier and recognize the crowd behaviours. In this research, a novel signature based on GLCM and its derived features is devised to describe the crowd behaviour within textures. The STT is firstly divided into patches with a grid. Then, the symmetric GLCM is calculated for each patch. Next, four derived features of GLCM including Contrast, Angular Second Moment, Entropy and Variance are modelled into the signature for the training and testing of the classifier. In the experiment, 5 classifiers including KNN, SVM, Naïve Bayes, DAC and random forest are adapted to classify the proposed signature and the conventional TAMURA texture feature on the UMN dataset and the simulated congested crowd. The result shows the average detection

accuracy reaches around 70%. The proposed signature outperforms TAMURA on UMN dataset, and the TAMURA has better performance on the simulated footage. Also, the combination of proposed signature and Naïve Bayes classifier has the best performance than others. The advantage of the proposed approach is the relatively high accuracy on the congestion and panic dispersing crowd abnormal behaviours. However, the computational time of this approach is long, since the process of STT and GLCM extraction is time consumptive. An alternative crowd behaviour detection approach is introduced as well.

In an early phase of this program, a change detection approach of panic dispersing is devised to achieve the fast detection of panic dispersing within the crowd. Instead of the local pattern, the trend of global motion flow is exploited as a feature to detect the sudden change among the crowd. In the research, the HS optical flow field is extracted and modelled into a feature for each frame. Next, the average feature value of the training data is calculated. In the testing phase, if the difference between current and average features is larger than the threshold for certain period of time, the anomaly is considered to exist. In the experiment, all dispersing behaviours are detected on all videos of UMN with minor deviations. The advantage of this approach is the fast processing speed. However, it is essentially a change detector, thus it can't classify the different types of crowd behaviours.

#### **8.1.4. Realistic Crowd Behaviour Synthesis and Prediction Methods**

In the research of crowd behaviour analysis, the acquisition of video data is an important problem. The benchmarking crowd video dataset is very limited and usually doesn't include the desired crowd behaviour. In order to obtain the required crowd video for analysis, the simulation tool could be utilized to simulate crowd with certain behaviours. In this research, a simulation approach involved with A-star path finding and enhanced social force model is proposed to generate crowd behaviour with visual realism. The behaviour model consists of three components including long-term path finding, local steering and interaction handling components. The long-term path finding is determined by the A-star algorithm. The local steering and interaction handling are

determined by the enhanced social force model. The enhanced social force model adapted the concepts of personal space, view perspective and relative velocity to improve the visual realism of the simulation. Furthermore, the group attraction force is proposed to simulate the behaviour of agents with same destination. In the experiment, seven crowd behaviour types in the proposed taxonomy are successfully synthesized. The key motion patterns and global appearance prove the visual realism level of the simulated crowd.

By transforming the formula of proposed behaviour model, the group attraction force and corresponding grouping center could be exploited to separate agents with different destinations when they are randomly distributed. In the research, agents are firstly detected and their instant motion flows are extracted. By calculating the actual affect force and estimating the repulsive and desired forces, the affected group attraction force could be obtained with the proposed formula. Therefore, the common grouping center will be used as a pattern to predict agents with same destination. In the experiment, the proposed approach is implemented on simulated crowds. Comparing to the spatial position, the grouping center exhibits the great performance on separating the crowd in the mixed state.

## 8.2. Future Work

Despite this research addressed some issues in the domain of crowd abnormal behaviour analysis, including the taxonomy of crowd behaviours, detection of congestion and panic dispersing, and crowd behaviour synthesis, several vital challenges remain to be tackled. This section covers some necessary directions to further expand the current research progress.

Many proposed approaches in this research involve the pedestrian detection techniques. However, in crowd with extremely high density, the frequent occlusion and the insufficient information make the low accuracy of pedestrian detection. Thus, the proposed approaches will fail in this situation. In order to address this issue, one

possible solution is to adapt the soft-NMS technique proposed in the research of Navaneeth *et al.* (2017) for the occlusion handling, along with the global pedestrian count estimation techniques such as the research of Antoni *et al.* (2008). By combining the global estimation and local occlusion handling with the conventional pedestrian detection technique, the challenge of pedestrian detection in crowd with extremely high density might be achieved.

In this research, the detection of congestion and panic dispersing has been addressed. However, according to the proposed taxonomy, several crowd behaviours are still not handled. In fact, no wide-range recognizing technique has been proposed in the domain of crowd behaviour analysis. Therefore, the proposed approach could be further expanded to handle more behaviour types. The ultimate goal is to devise a universal recognition technique which is capable of handling wide-range of crowd behaviours.

The CV based crowd abnormal behaviour analysis exploits only video data. Some researches use network data collected from social media to predict the possible behaviour of the crowd within certain region. In the research of Michalis *et al.* (2017), the video and audio data are simultaneously adapted for the crowd behaviour analysis. Therefore, exploiting data from multiple sources including video, audio and network information could be another possible path to improve the performance of proposed analysis approaches.

---

## References

---

- Adelson E.H. and Bergen J.R. (1985). Spatio-temporal energy models for the perception of motion. *Journal of the Optical Society of America*, **2**(2):284-299.
- Alex P. P. (1984). Fractal-Based Description of Natural Scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence 1984*, **6**(1):661-674.
- Alexandre A., Kratharth G., Vignesh R., Alexandre R., Li F. and Silvio S. (2016). Social LSTM: Human Trajectory Prediction in Crowded Spaces. *The IEEE Conference on Computer Vision and Pattern Recognition 2016 (CVPR16)*. Las Vegas, US.
- Ali S. and Shah M. (2007). A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. *The IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR07)*. Minneapolis, US.
- Antoni B. C., Zhang S., John L. and Nuno V. (2008). Privacy Preserving Crowd Monitoring: Counting People without People Models or Tracking. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage.
- Barbara K. and Christian B. (2012). Loveparade 2010: Automatic video analysis of a crowd disaster. *Computer Vision and Image Understanding*, **116**(1):307–319
- Baum L. E. and Petrie T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, **37**(6):1554–1563.
- Beauchemin S. S. and Barron J. L. (1995). The computation of optical flow. *ACM Computing Surveys (CSUR)*, **27**(3):433-466.
- Berlonghi A. E. (1995). Understanding and planning for different spectator crowds. *Safety Science*, **18**(4):239–247
- Blue V.J. and Adler J.L. (1998). Emergent fundamental pedestrian flows from cellular automata microsimulation. *Transportation Research Record*, **1644**(1):29-36.
- Bo W., Mao Y., Xue L., Fengjuan Z. and Jian D. (2012). Abnormal crowd behaviour detection using high-frequency and spatio-temporal features. *Machine Vision and Applications*, **23**(3):501-511.
- Bobick A.F. and Davis J.W. (2001). The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(3):257-267.
- Bolles R. C., Baker H. H. and Marimont D. H. (1987). Epipolar-plane image analysis: an approach to determining structure from motion. *International Journal of Computer Vision*, **1**(1):7–55.

- 
- Bottou L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings of COMPSTAT'2010*, **1**(1):177-186
- Chenney, S. (2004). Flow Tiles. *Eurographics Proceedings of Symposium on Computer Animation*, Grenoble, France.
- Christian S., Wei L. *et al.* (2015). Going Deeper With Convolutions. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, Massachusetts, USA.
- Cortes C. and Vapnik V. N. (1995). Support-vector networks. *Machine Learning*, **20**(3):273–297.
- Craig W. R. (1987). Flocks, Herds, and Schools: A Distributed Behavioural Model. *Computer Graphics*, **21**(4): 25-34.
- Cui X., Liu Q. and Gao M. (2011). Dimitris N. M., Abnormal Detection Using Interaction Energy Potentials. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA.
- Dalal N. and Triggs B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA.
- Deepak K. P. and Meher S. (2015). Hierarchical background subtraction algorithm using Gabor filter. *Proceedings of IEEE International Conference on Electronics, Computing and Communication Technologies*, Bangalore, India.
- Diederik P. K., Jimmy B. (2015). Adam: A Method for Stochastic Optimization. *The 3rd International Conference for Learning Representations*, San Diego, USA.
- Du T., Lubomir B., Rob F., Lorenzo T. and Manohar P. (2015). Learning Spatiotemporal Features With 3D Convolutional Networks. *The IEEE International Conference on Computer Vision (ICCV 2015)*, Santiago, Chile.
- Ephraim Y. and Malah D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **32**(6):1109-1121.
- Ernesto L. A., Scott B. and Robert B. F. (2006). Hidden Markov Models for Optical Flow Analysis in Crowds. *18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, China.
- Feifei L. and Perona P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA.

---

Florent P. and Christopher D. (2007). Fisher Kernels on Visual Vocabularies for Image Categorization. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA.

Ghosh S. and Mallett R.L. (1994). Voronoi cell finite elements. *Computers and Structures*, **50**(1): 33-46.

Greg P., Ramin Z., Justin M. (1996). Comparing images using color coherence vectors. *MULTIMEDIA '96 Proceedings of the fourth ACM international conference on Multimedia*, Boston, Massachusetts, USA.

Hamidreza R., Javad H. and Hossein M. (2016). Crowd behaviour representation: an attribute-based approach. *SpringerPlus*, **5**(1):1179.

Haralick R. M., Shanmugam K. and Dinstein I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, **3**(6):610–621.

Hart P. E., Nilsson N. J. and Raphael B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics SSC4*, **4**(2):100–107.

He X., Zemel R.S. and Carreira-Perpiñán M.Á. (2004). Multiscale conditional random fields for image labeling. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, USA.

Helbing D., and Peter M. (1995). Social force model for pedestrian dynamics. *Physical review*, **51**(5):4282.

Helbing D., Farkas I., Molnar P. and Vicsek T. (2002). Simulation of pedestrian crowds in normal and evacuation situations. *Pedestrian and evacuation dynamics*, Duisbueg, Germany.

Horn B.K.P. and Schunck B.G. (1981). Determining optical flow. *Artificial Intelligence*, **17**(1): 185–203.

Howard E. and Haluk D. (1987). Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*1987, **9**(1):39-55.

Hui Y., Mingjing L., HongJiang Z. and Jufu F. (2002). Color texture moments for content-based image retrieval. *Proceedings 2002 International Conference on Image Processing*, Rochester, NY, USA.

Ivan L. (2005). On Space-Time Interest Points. *International Journal of Computer Vision*, **64**(2–3):107–123.

Ivan L., Marcin M., Cordelia S. and Benjamin R. (2008). Learning realistic human actions from movies. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, USA.

- 
- Jing W. and Zhijie X. (2016). Spatio-temporal texture modelling for real-time crowd anomaly detection. *Computer Vision and Image Understanding*, **144**(1):177-187
- Junqi J., Kun F. and Changshui Z. (2014). Traffic Sign Recognition With Hinge Loss Trained Convolutional Neural Networks. *IEEE Transactions on Intelligent Transportation Systems*, **15**(5):1991-2000
- Kai H., Zhenzhen Z. *et al.* (2018). Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function. *Neurocomputing*, **309**(2):179-191
- Kai K. and Xiaogang W. (2014). Fully Convolutional Neural Networks for Crowd Segmentation. *ArXiv 2014*.
- Karen S. and Andrew Z. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *ArXiv 2014*.
- Karen S. and Andrew Z. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv 2014*.
- Ketkar N. (2017). Introduction to PyTorch. *Deep Learning with Python*. Apress, Berkeley, CA.
- Kishore K. R., and Mubarak S. (2012). Recognizing 50 Human Action Categories of Web Videos. *Machine Vision and Applications Journal (MVAP)*.
- Koralov L., and Sinai Y.G. (2012). Gibbs Random Fields. *In Theory of Probability and Random Processes* (pp. 341-346). Springer, Berlin, Heidelberg.
- Kuehne H., Jhuang H., Garrote E., Poggio T., and Serre T. (2011). HMDB: A Large Video Database for Human Motion Recognition. *In the 13th International Conference on Computer Vision*. Barcelona, Spain.
- Kuhne G, Richter S. and Beier M. (2001). Motion-based segmentation and contour-based classification of video objects. *Proceedings of the 9th ACM international conference on Multimedia*. Ottawa, Canada.
- Lazebnik S., Schmid C. and Ponce J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. New York, USA
- Lazebnik S., Torralba A., Fei-Fei L., Lowe D. and Szurka C. (2012). Tutorials of Bag-of-Words models. Web site: [https://cs.nyu.edu/~fergus/teaching/vision\\_2012/9\\_BoW.pdf](https://cs.nyu.edu/~fergus/teaching/vision_2012/9_BoW.pdf)
- Lee J., Jin R., and Jain K. A. (2012). Image Retrieval in Forensics: Tattoo Image Database Application. *IEEE Multimedia*, **19**(1):40-49.

- 
- Lei Q., Zhongwei C. Qingming H. and Junbiao P. (2012). Interactive event detection in crowd scenes. *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*, New York, NY, USA.
- Li S. Z. (1994). Markov random field models in computer vision. *European Conference on Computer Vision (ECCV'94)*, Berlin, Heidelberg.
- Lowe D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision*. Kerkyra, Greece.
- Lucas B. and Kanade T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. *Proceedings of Imaging Understanding Workshop*. Vancouver, Canada.
- Ma D., Wang Q. and Yuan Y. (2014). Anomaly Detection in Crowd Scene via Online Learning. *Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS14)*, Xiamen, China.
- Martin A., Paul B. *et al.* (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, Savannah, GA, USA.
- Mehran R, Moore B. E., and Shah M. (2010). A streakline representation of flow in crowded scenes. *European Conference on Computer Vision (ECCV 2010)*, Berlin, Heidelberg.
- Mehran R. (2009). Abnormal crowd behaviour detection using social force model. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA.
- Michalis V., Christophoros N. and Ioannis A. K. (2017). Identifying Human Behaviours Using Synchronized Audio-Visual Cues. *IEEE Transactions on Affective Computing*, **8**(1):54-66
- Mikel R., Josef S., Ivan L. and Jean-Yves A. (2011). Data-driven Crowd Analysis in Videos. *13th International Conference on Computer Vision (ICCV 2011)*, Barcelona, Spain.
- Momboisse R. (1967). Riots Revolts, and Insurrection. *Springfield, Ill. Charles Thomas*.
- Navaneeth B., Bharat S., Rama C. and Larry S. D. (2017). Improving Object Detection With One Line of Code. *International Conference on Computer Vision 2017*. Venice, Italy.
- Ngo C. W., Pong T. C. and Chin R. T. (1999). Detection of gradual transitions through temporal slice analysis. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Fort Collins, USA.
- Nguyen N.A.T., Zucker J.D., Nguyen H.D. and Drogoul A. (2011). A Hybrid macro-micro pedestrians evacuation model to speed up simulation in road networks. *Advanced Agent Technology. AAMAS 2011. Lecture Notes in Computer Science*, Taipei, Taiwan.

---

Novak C.L., Shafer S.A. (1992). Anatomy of a color histogram. *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Champaign, IL, USA.

Ojala T. Pietikainen M. and Maenpaa T. (2002). Multiresolution Grey-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(7):971-987.

Park S. I., Cao Y. and Quek F. (2011). Large Scale Crowd Simulation using A Hybrid Agent Model. *The fourth International Conference Motion in Games*, Edinburgh, UK.

Rabaud V., Cottrell G., Belongie S. and Dollar P. (2005). Behaviour recognition via sparse spatio-temporal features. *Proceedings of 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, Beijing, China.

Raghavendra R., Del B. A., Cristani M. and Murino V. (2011). Abnormal Crowd Behaviour Detection by Social Force Optimization. *Salah A.A., Lepri B. (eds) Human Behaviour Understanding. HBU 2011. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg.

Ranjan R. K. and Agrawal A. (2016). Video summary based on F-sift, Tamura textural and middle level semantic feature. *Procedia Computer Science*, **89**(1):870–876.

Rathinavel S. and Arumugam S. (2011). Full Shoe Print Recognition based on Pass Band DCT and Partial Shoe Print Identification using Overlapped Block Method for Degraded Images. *International Journal of Computer Applications*, **26**(8):3126-4301

Reicher S. and Alan E. K. (2000). Encyclopedia of psychology. *Washington, D.C.: American Psychological Association*.

Reynolds C. (1987). Flocks, Herds, and Schools: A Distributed Behavioural Model. *Computer Graphics*, **21**(4):25-34.

Richard O. D. and Peter E. H. (1972). Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, **15**(1):11-15.

SangHyun C. and HangBong K. (2014). Abnormal behaviour detection using hybrid agents in crowded scenes. *Pattern Recognition Letters*, **44**(1):64-70.

Saxena S., Brémond F., Thonnat M. and Ma R. (2008). Crowd Behaviour Recognition for Video Surveillance. *Advanced Concepts for Intelligent Vision Systems*, **5259**(1):970-981.

Sean R. E. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, **6**(3):361-365.

Sejdić E., Djurović I. and Jiang J. (2009). Time-frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, **19**(1):153-183.

- 
- Seung-Hwan B. and Kuk-Jin Y. (2014). Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA.
- Sevilla-Lara L., Liao Y., Güney F., Jampani V., Geiger A. and Black M.J. (2019). "On the Integration of Optical Flow and Action Recognition". German Conference on Pattern Recognition 2018. Stuttgart, Germany.
- Sewall J., Wilkie D. and Lin M.C. (2011). Interactive hybrid simulation of large-scale traffic. *Proceedings of the 2011 SIGGRAPH Asia Conference*, Hong Kong, China.
- Shannon C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**(3): 379–423.
- Shen L. and Bai L. (2004). Gabor feature based face recognition using kernel methods. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, South Korea.
- Sivic J. (2009). Efficient visual search of videos cast as text retrieval. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, **31**(4):591–605.
- Solmaz B., Moore B. E. and Shah M. (2012). Identifying behaviours in crowd scenes using stability analysis for dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(10):2064–2070.
- Somayeh D., Robert W. and Anna-Katharina L. (2008). Towards a taxonomy of movement patterns. *Information Visualization*, **7**(3-4):240 – 252
- Stuart C., Fotis L. and Chrisina J. (2015). Perceived Realism of Crowd Behaviour with Social Forces. *19th International Conference on Information Visualisation*, Barcelona, Spain.
- Tamura H., Mori S. and Yamawaki T. (1978). Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man, and Cybernetics*, **8**(6):460-473.
- Tara N. S., Oriol V., Andrew S. and Hasim S. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, Australia.
- Tissera P.C., Printista A.M. and Luque E. (2012). A Hybrid Simulation Model to Test Behaviour Designs in an Emergency Evacuation. *Proceedings of the International Conference on Computational Science*, **9**(1):266-275.
- Vikas R., Conrad S. and Brian C. L. (2011). Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. *Computer Vision and Pattern Recognition 2011 Workshop*, Colorado Springs, CO, USA.

- 
- Venkatesh S. and Zhu C. (2012). Video Anomaly Detection Based on Local Statistical Aggregates. *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA*.
- Viola P. and Jones M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA*.
- Wang H., Kläser A., Schmid C. Chenglin L. (2013). Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision*, **103**(1): 60-79.
- Wang H. and Schmid C. (2013). Action Recognition with Improved Trajectories. *IEEE International Conference on Computer Vision (ICCV)*, Sydney, NSW, Australia.
- Wilhelm B. and Mark J. B. (2013). Fourier Shape Descriptors. *Principles of Digital Image Processing*, **1**(1):169-227.
- Xinyi C., Qingshan L., Mingchen G. and Dimitris N. M. (2011). Abnormal Detection Using Interaction Energy Potentials. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA
- Xiong M., Cai W. and Suiping Z. (2009). A case study of multi-resolution modeling for crowd simulation. *Proceedings of the 2009 Spring Simulation Multiconference*, San Diego, California.
- Xuxin Gu, Jinrong Cui, Qi Zhu (2014). Abnormal crowd behaviour detection by using the particle entropy. *Optik*, **125**(14):3428-3433.
- Yang B., Sang X. and Xing S. (2016). A-star algorithm based path planning for the glasses-free three-dimensional display system. *Proceedings of SPIE - The International Society for Optical Engineering*, Beijing, China.
- Yang, J., Bai, Y., Lin, F. *et al.* (2018). A novel electrocardiogram arrhythmia classification method based on stacked sparse auto-encoders and softmax regression. *International Journal of Machine Learning and Cybernetics*, **9**(10):1733-1740.
- Yangqing J., Evan S., Jeff D., Sergey K., Jonathan L., Ross G., Sergio G. and Trevor D. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *Proceedings of the 22nd ACM international conference on Multimedia*, Orlando, Florida, USA
- Zhou D. X., Zhang H. and Ray N. (2008). Texture based background subtraction. *Proceedings of IEEE International Conference on Information and Automation*, Changsha, China.