

VISION-BASED CROWD CONGESTION MANAGEMENT IN TRANSPORT HUBS

YAN LI

Submitted in total fulfillment of the requirements of
the degree of Doctor of Philosophy

November 2019

Department of Infrastructure Engineering
The University of Melbourne

Yan Li: *Vision-based crowd congestion management in transport hubs*

© November 2019

SUPERVISORS:

Prof Majid Sarvi

Dr Kourosh Khoshelham

LOCATION:

Melbourne, Australia

ABSTRACT

Congestion in transport hubs has been a big issue in many population booming cities as the current transport systems are not able to accommodate the increasing surge in travel demand caused by ongoing urbanisation and growth of the world population. The ubiquitous Closed-Circuit Television (CCTV) cameras in transport hubs provide a means for automatic crowd surveillance by utilising accurate and robust image-based crowd analysis models. However, this is an active field of research due to the challenges of occlusion among a large number of individuals and environmental changes, such as lighting fluctuations and changes in context over time.

The research presented in this thesis aims to understand and manage the station congestion from the macro and micro perspectives by presenting a framework enabling individual origin-destination estimation and crowd congestion map generation. The main contributions of the study can be outlined as:

- development of spatial-temporal deep learning models for macro-scale crowd density map generation, and micro-scale pedestrian Origin-Destination (O-D) estimation based on person re-identification in the internet of cameras.
- a multi-modal framework for crowd analysis under emergency scenarios, especially low visibility situation, using RGB-D cameras.
- linking models and application by developing a web-GIS platform for real-time processing and visualization, thereby demonstrating the applicability of the research in crowd congestion management.

This research presents distinct interdisciplinary components with relevance to crowd dynamics as well as image processing and with practical implications for smart station management.

KEYWORDS: crowd congestion, visual congestion analysis, level of service, pedestrian origin-destination, congestion prediction, deep learning, convolutional neural network, long short-term memory

DECLARATION

This is to certify that:

1. the thesis comprises only my original work towards the PhD except where indicated in the Preface,
2. due acknowledgment has been made in the text to all other material used,
3. the thesis is fewer than 100 000 words in length, exclusive of tables, maps, and bibliographies.

Melbourne, Australia, November 2019

Yan Li

PUBLICATIONS

This thesis is based on published work from my PhD research during candidature. Some ideas and figures have appeared previously in the following publications:

PUBLISHED ARTICLES

- Li, Y., Khoshelham, K., Sarvi, M., & Haghani, M.. Direct generation of level of service maps from images using convolutional and long short-term memory networks. *Journal of Intelligent Transportation Systems*, **2019**, 1-9.
- Li, Y., Sarvi, M., Khoshelham, K., Haghani, M.. Multi-view crowd congestion map generation based on ensemble learning, In: *Transportation Research Board 98th Annual Meeting Transportation Research Board*, **2019**, (No. 19-04444).
- Li, Y., Sarvi, M., Khoshelham, K.. Pedestrian origin-destination estimation in emergency scenarios. In: *9th International Conference on Fire Science and Fire Protection Engineering (ICFSFPE 2019)*, **2019**.
- Li, Y., Sarvi, M., Khoshelham, K., & Haghani, M.. Real-Time Level-of-Service Maps Generation from CCTV Videos. In: *Transportation Research Board 97th Annual Meeting Transportation Research Board*, **2018**, (No. 18-06670).
- Li, Y., Sarvi, M., Khoshelham, K., & Haghani, M.. Predicting traffic congestion maps using convolutional long-short term memory. In : *International Symposium of Transport Simulation (ISTS' 18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS' 18)*, **2018**.

ARTICLES UNDER REVIEW

- Li, Y., Sarvi, M., Khoshelham, K., Haghani, M.. Multi-view crowd congestion monitoring system based on an ensemble of CNN classifiers. *Journal of Intelligent Transportation Systems*. (In revision).
- Li, Y., Sarvi, M., Khoshelham, K.. Predicting traffic congestion using convolutional long short-term memory networks. *Transportation Research Part C: Emerging Technologies*. (In revision)

- Li, Y., Sarvi, M., Khoshelham, K.. Pedestrian origin-destination estimation based on multi-camera person re-identification. *IEEE Internet of Things*. (In revision)

ACKNOWLEDGMENTS

The three and half years working on this thesis is like running a marathon, which is my favourite sport. It takes a lot of courage and tenacity to keep going till the end, which cannot be achieved without the contributions of many people.

First and foremost, I would like to express my gratitude to my supervisors for guiding me like the pacemaker in a marathon and helping me to realize my time plan at an achievable pace. My main supervisor Prof. Majid Sarvi, as the pioneer and leader of Australian Integrated Multimodal EcoSystem (AIMES) project provided me many excellent ideas and comments. Without his guidance, encouragement and support, this thesis would not have existed. My sincere thanks are extended to my co-supervisor Dr. Kourosh Khoshelham. Kourosh led me to overcome the obstacles in scientific thinking and writing on the way to become a researcher. I would also like to thank Dr. Milad Haghani for all the feedback and suggestions, who also shared my daily trials and triumphs, successes and frustrations, thank you for encouragement and wise words when needed. I also express my gratitude to Prof. Abbas Rajabifard, Associate Prof. Russell Thompson and Associate Prof. Yongping Wei for accepting me into the University of Melbourne and kindly giving me the freedom to choose the topic I am happy with.

My research would certainly not have been possible without Mr. Fred Wang (Public Transport Victoria) and Mr. Tony Fusco (Metro Trains) for their tremendous help in data and ideas on the topic.

Writing a PhD thesis can be a lonely experience, I am fortunate to be able to work with many great teams and members, where I always get inspired on both research and life. I would like to thank all the colleagues in our amazing Transport Engineering Group, Dr. Zahra Shahhoseini, Dr. Zay Maung Maung, Mr. Abdullah Alhawsawi, Ms. Azadeh Emami, Mr. Mohammad Mobin Yazdani, colleagues from our Vision and Positioning Lab, Dr. Fuqiang Gu, Dr. Milad Ramezani, Mr. Debaditya Acharya, Ms. Ha Thi Thu Tran, Mr. Hanxian He, Mr. Marko Radanovic, Mr. Hang Zhao, as well as colleagues in our department of Infrastructure Engineer-

ing, Dr. Haifeng Zhao, for support and discussion on related research questions and sharing with experiences.

I would also like to thank the past and present fellow researchers at Center for Disaster Management and Public Safety (CDMPS) and Centre for Spatial Data Infrastructure and Land Administration (CSDILA) for your support and friendship during my candidature, Prof. Greg Foliente, Dr. Mohsen Kalantari, Dr. Yiqun Chen, Dr. Soheil Sabri, Dr. Arash Kavani, Dr. Mitko Alexanderov, Dr. Hewage Loshaka Kumara Perera, Dr. Cheng Cheng, Mr. Kuo-Chih Hung, Mr. Jude Shalitha Perera, Ms. Jing Ren, Ms. Anne Wang, Ms. Yanyu Wang, Ms. Yang Yang, Ms. Junbi Zhou, Ms. Chaojie Guo.

Thanks are also to the visiting staff to the Department who provided their invaluable advice and research perspective, especially Prof. Armin Seyfried, Dr. Maik Boltes (Juelich Supercomputing Centre), Associate Prof. Daichi Yanagisawa (University of Tokyo), Prof. Bo Fan (Shanghai Jiaotong University), Prof. Shen Ying (Wuhan University), Prof. Hao Sheng (Beihang University), Prof. Hongsheng Li (Chinese Academy of Sciences), Associate Prof. Xin Li (China University of Mining and Technology), Mr Jiangyu Wang (Southwest Jiaotong University). My thanks also go to Prof. Jian Ma, Prof. Peng Lin (Southwest Jiaotong University) and Ms Jinglin He (Wuhan University) for the hospitality during my visit to the conferences.

Finally, I would like to express my deepest gratitude to my parents, my family for their patience, support and encouragement, especially during the most stressful period of producing this thesis. Although the PhD marathon would end soon, the marathon of our life would continue. Thank you for love!

CONTENTS

I	TOWARDS VISUAL CONGESTION ANALYSIS	1
1	INTRODUCTION	3
1.1	Motivation	3
1.1.1	Congestion in transport hubs issue	3
1.1.2	CCTV camera surveillance	3
1.1.3	Visual crowd analysis	4
1.1.4	Lack of congestion vision	5
1.2	Research questions	6
1.3	Research aim and scope	7
1.4	Major contributions	7
1.5	Thesis structure	8
2	LITERATURE REVIEW	11
2.1	Crowd congestion indicators	11
2.2	Single view crowd congestion map generation . . .	13
2.3	Multi-view crowd congestion map generation . . .	15
2.4	Congestion prediction	16
2.5	Pedestrian Origin-destination estimation	18
2.6	Crowd analysis using RGB-D cameras	21
II	MACRO-SCALE CROWD CONGESTION MAP GENERATION	29
3	SINGLE VIEW LEVEL OF SERVICE MAP GENERATION	31
3.1	Introduction	31
3.2	Overview of the approach	32
3.2.1	Defining the region of interest	32
3.2.2	Creating grids	33
3.2.3	Feature extraction and classification	34
3.2.4	Refining level of service labels	35
3.2.5	Web-based visualization of level of service maps	36
3.3	Experimental evaluation	37
3.3.1	Data sets	37
3.3.2	LOS classification accuracy	37
3.3.3	Computational efficiency	37
3.3.4	Temporal analysis of LOS	38
3.4	Summary and conclusions	40
4	MULTI-VIEW LEVEL OF SERVICE MAP GENERATION	43

4.1	Introduction	43
4.2	Multi-camera fusion framework	44
4.2.1	Defining sub-images corresponding to the grid cells	45
4.2.2	Single view level of service classification . .	46
4.2.3	Merging level of service at different views .	47
4.2.4	Level of service map visualization	49
4.3	Experiments and findings	50
4.3.1	Datasets	51
4.3.2	Comparison of ensemble rules	53
4.3.3	Effect of number of views	56
4.3.4	Computational efficiency	57
4.4	Conclusions	58
5	PREDICTING TRAFFIC CONGESTION USING CONVOLU- TIONAL LONG SHORT-TERM MEMORY NETWORKS	61
5.1	Introduction	61
5.2	Methodology	62
5.2.1	Problem statement	62
5.2.2	Network architecture	63
5.3	Experiments	65
5.3.1	Datasets	65
5.3.2	Data pre-processing	67
5.3.3	Experiment settings	67
5.3.4	Evaluation metrics	68
5.3.5	Experiment results	68
5.3.6	Application of congestion length calculation	70
5.4	Conclusion and future work	72
III	MICRO-SCALE PEDESTRIAN ORIGIN-DESTINATION ES- TIMATION	75
6	PEDESTRIAN ORIGIN-DESTINATION ESTIMATION BASED ON MULTI-CAMERA PERSON RE-IDENTIFICATION	77
6.1	Introduction	77
6.2	Methodology	78
6.2.1	Frame extraction and synchronization . . .	79
6.2.2	Pedestrian detection and generation of gallery images	79
6.2.3	Pedestrian re-identification	80
6.2.4	O-D flow generation	82
6.3	Experiments and results	84
6.3.1	Dataset	84

6.3.2	Evaluation	84
6.3.3	O-D flow estimation	86
6.4	Conclusions and future perspectives	87
7	PEDESTRIAN ORIGIN-DESTINATION ANALYSIS USING DEPTH CAMERAS	91
7.1	Introduction	91
7.2	Methodology	92
7.3	Datasets	94
7.3.1	Camera setup	94
7.3.2	Scenarios	94
7.3.3	Ground-truths	97
7.4	Experiments and results	97
7.5	Conclusions and future work	99
IV	CONCLUSION AND FUTURE WORK	101
8	CONCLUSION	103
8.1	Research contributions	103
8.2	Major results and findings	103
8.2.1	Macro-scale congestion	103
8.2.2	Micro-scale congestion	105
9	LIMITATIONS AND FUTURE WORK	107
9.1	Internet of visual sensors	107
9.2	From traffic to crowd congestion prediction	107
9.3	Multi-modality origin-destination estimation	108
	BIBLIOGRAPHY	111

LIST OF FIGURES

Figure 1	Station congestion can be seen from the daily ridership number of big cities.	4
Figure 2	The main tasks of visual crowd analysis . .	5
Figure 3	Research framework	7
Figure 4	Overview of the approach	32
Figure 5	Geometric transformation of the region of interest into a rectangular image	33
Figure 6	Structure of deep learning network for LOS classification	35
Figure 7	A Cesium-based interface is developed to visualize interactive LOS maps. The 3 dimensional model of Flinders Street Station is presented to better locate the congestion positions.	36
Figure 8	LSTM refinement method (green line) which considers temporal relations is more close to the ground-truth (blue line) than non-temporal method (red line), which is more fluctuate around ground-truth.	38
Figure 9	level of service map update in 5-minute intervals	39
Figure 10	Temporal variation of level of service in each grid	40
Figure 11	System architecture of multi-view crowd congestion map fusion system	45
Figure 12	An illustration of define sub-images corresponding to the grid cells.	46
Figure 13	Single view level of service classification based on a Convolutional Neural Network and Support Vector Machine. LOS A to F is equal to 0-1, 2, 3, 4, 5-8, 9 or more pedestrians per grid with a grid cell size of $2m \times 2m$, respectively.	47
Figure 14	The overall LOS maps is visualized in a web-based interface. The fields of view of each camera are outlined by colored lines.	50

Figure 15	The camera views are projected to the ground view plane after computing the homograph matrix for all camera views.	52
Figure 16	Comparisons of ensemble learning techniques. As seen, ensemble learning techniques can help improve classification accuracy. Algebraic based methods show better performance than other techniques.	54
Figure 17	Performance of algebraic sum rule. PETS2009 dataset shows better performance in low LOS (from A to D), and Mecca dataset has higher accuracy in high LOS (from C to F).	55
Figure 18	Combinations of cameras are evaluated on PETS2009 dataset using algebraic sum rule. Two camera views with high resolution and no occlusion can achieve optimized performance.	57
Figure 19	Loading test, which measures the maximum number of cameras executed on this system. In order to achieve a map refresh rate of 2 seconds, maximum 48 cameras views can be run at the same time under the test environment.	58
Figure 20	Input data formulation	63
Figure 21	ConvLSTM model based prediction structure	65
Figure 22	Datasets:(a) Freeway I80 in San Francisco. (b) The minor roads in United Kingdom.	66
Figure 23	Accuracy of each predicted frame before and after enhancement.(a) PEMS180 dataset. (b)UKMINR dataset.	69
Figure 24	Some prediction examples, from top to bottom: input frames; ground truth images and prediction images before enhancement, prediction images after enhancement. (a) PEMS180 dataset. (b)UKMINR dataset.	71
Figure 25	Congestion length of one road, the total of red and brown parts	72
Figure 26	Congestion length by time.	73
Figure 27	The proposed framework for pedestrian origin-destination estimation.	78

Figure 28	The idea of calculating percentage of visible body part.	81
Figure 29	Pairwise siamese model takes two feature vectors as input while triplet model takes three feature vectors as inputs.	83
Figure 30	Trajectory data structure.	83
Figure 31	The camera map of DukeMTMC-reID dataset	84
Figure 32	CMC curve of image-based methods under different congestion levels.	89
Figure 33	Pedestrian O-D flow in the first half hour (a) and the second half (b) of the video. . .	90
Figure 34	Overview of the framework. The color, depth and infrared images are processed separately and fused in the last fully connected layer to generate a fused feature for classification.	93
Figure 35	Evacuation map	97
Figure 36	Correct O-D detections by LOS in good and poor lighting conditions based on RGB-D-I-CNN method. There are 16 O-D routes per LOS.	98

LIST OF TABLES

Table 1	Crowd congestion indicators	12
Table 2	List of 3D consumer cameras	25
Table 3	RGBD datasets for pedestrian recognition .	26
Table 4	Datasets for behaviour analysis	27
Table 5	Level of service: the most widely accepted congestion index (Source: Highway capac- ity manual (National Research Council (U.S.) (2010)))	34
Table 6	Number of samples and accuracy for each level	38
Table 7	Average processing time for each frame . .	39
Table 8	An example of deciding the final level of service based on several combination rules	50
Table 9	Confusion matrix using algebraic-sum based method.	56
Table 10	The prediction accuracy at different time intervals on PEMS18o dataset	68
Table 11	The prediction accuracy at different time intervals on UKMINR dataset	70
Table 12	The relationships between LOS and the per- centage of visible body parts.	81
Table 13	Model performance using cross entropy (top- 1 rate).	86
Table 14	Model performance using triplet loss (top- 1 rate).	86
Table 15	Origin-destination (O-D) table	87
Table 16	Snapshots at classroom doorway (poor light- ing). The color cameras perform very poorly in this situation.	95
Table 17	Snapshots at classroom doorway (good light- ing)	96
Table 18	Rank-1 accuracy in good lighting condition (%)	99
Table 19	Rank-1 accuracy in poor lighting condition (%)	99

ACRONYMS

CCTV Closed-Circuit Television

GIS Geographic information system

LOS Level of service

DOC Degree of crowding

FD Fundamental diagram

NFD Network fundamental diagram

O-D Origin-destination

DNN Deep neural network

CNN Convolutional neural network

RNN Recurrent neural network

LSTM Long short-term memory

ConvLSTM Convolutional long short-term memory

Part I

TOWARDS VISUAL CONGESTION ANALYSIS

This part explains the background and motivation of this thesis. Based on a review of current visual congestion analysis techniques and applications, Chapter [2](#) discusses the concept of crowd congestion, and the up-to-date status of visual congestion analysis in both application and technology domains, and the challenges of visual congestion analysis that hinder its further promotion.

INTRODUCTION

1.1 MOTIVATION

1.1.1 *Congestion in transport hubs issue*

Crowd congestion has been a big issue in transport hubs of many population booming cities. With over 68% of the world's population expected to be living in cities by 2050 ¹, supporting citizen's mobility within the urban environment is a priority for municipalities worldwide. However, present transit systems in many busiest stations, including London, Hong Kong, and Melbourne, are not able to accommodate increasing loads, especially at peak commuting hours when the system has to cope with a temporary surge in demand. Figure 1 lists the daily ridership of the transport hubs in some of the world's big cities. It can be seen that congestion creates incredibly high levels of discomfort. A serious consequence of congestion is that crowd accidents might happen, such as people falling off the platform or being pushed onto the tracks. In history, a total of 374 people died and hundreds were injured in the stampedes happening at train stations around the world ². Crowd safety problem caused by congestion cannot be ignored.

1.1.2 *CCTV camera surveillance*

CCTV cameras provide a means for transport station managers to monitor the crowd for safety and operational purpose. However, as CCTV becomes ubiquitous, it becomes impossible for humans to monitor all available viewpoints. According to Victoria Police data, more than 2300 cameras have been installed in the metropolitan stations ³. Meanwhile, China launches Skynet project, with

¹ 68% of the world population projected to live in urban areas by 2050, says UN. <https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html>

² List of human stampedes and crushes. https://en.wikipedia.org/wiki/List_of_human_stampedes_and_crushes

³ Better Cameras To Make Train Travel Safer. <https://www.premier.vic.gov.au/better-cameras-to-make-train-travel-safer/>



(a) Wuhan: 3.1 million



(b) Melbourne: 0.66 million



(c) Sydney: 0.98 million



(d) Tokyo: 6.8 million



(e) Hong Kong: 4.9 million



(f) New York: 5.6 million

Figure 1: Station congestion can be seen from the daily ridership number of big cities.

an estimated 170 million cameras installed nationwide, and the number of cameras will increase to 400 million by 2020 ⁴. Therefore, automatic crowd monitoring from CCTV footage has aroused great interests from researchers, and developed into a research area called *visual crowd analysis*.

1.1.3 Visual crowd analysis

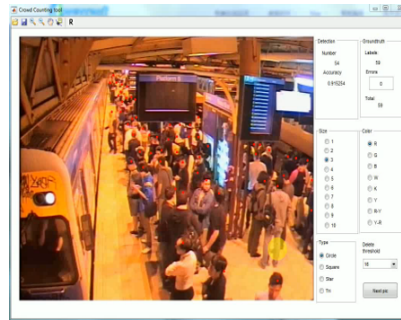
Visual crowd analysis is the use of visual sensors to conduct crowd analysis tasks. In the literature [Tripathi, Singh, and Vishwakarma, 2019; Gayathri, Aparna, and Verma, 2017; Grant and Flynn, 2017; Kaiser et al., 2017; Lamba and Nain, 2017; Yogameena and Nagananthini, 2017; Zitouni et al., 2016; Ibrahim et al., 2016], the

⁴ In Your Face: China's all-seeing state. <https://www.bbc.com/news/av/world-asia-china-42248056/in-your-face-china-s-all-seeing-state>

main tasks of visual crowd analysis include crowd detection and tracking, crowd counting, and crowd behavior analysis. Figure 2 shows examples of crowd analysis tasks. Crowd detection and tracking is to detect the pedestrians from the images and then track the individuals detected. As shown in Figure 2a, colored shapes outline detected pedestrians and each identity is assigned to one color. During the tracking process, each tracked identity would retain the same color, so that the trajectory of each pedestrian can be extracted, and the traffic parameters, such as speed, direction can be calculated based on the trajectory. Crowd counting is to count the number of people in camera view or defined region of interest. As shown in Figure 2b, the number of passengers on the platform is displayed in the interface. Crowd anomaly detection is to detect abnormal activities in a crowd, for example, a person is moving in the opposite direction of the crowd, as shown in Figure 2c.



(a) Crowd detection and tracking



(b) Crowd counting



(c) Abnormality detection

Figure 2: The main tasks of visual crowd analysis

1.1.4 Lack of congestion vision

Despite the numerous studies that have been made in the field of visual crowd analysis, the research focuses mainly on computer

vision techniques, collision analysis, or evacuation. Little research has explored the issue of *congestion*, either the vision of congestion or the vision for congestion.

The *vision of congestion* is an insight into congestion. Congestion describes the phenomena that the distance between individuals gets smaller and the behavior of each individual will be affected by the occupancy of others. Although pedestrian counting and density can be applied to estimate the level of congestion in human crowds, they are not the only relevant property making a particular crowd more congested (or dangerous) than another one. Pedestrian flow can be also used to estimate the amount of motion, but how to relate flow to congestion is unclear. In other words, the definition of congestion is rather vague and requires further research in a way that can be accurately quantified.

The *vision (eye) for congestion* is the visual sensor for congestion analysis. CCTV camera is the most widely used visual sensor for crowd surveillance, there are still many unsolved problems due to the challenges of occlusion among a large number of individuals, as well as environmental changes, such as lighting fluctuations and changes in context over time. The improvements to current vision-based models and the possibility of using other sensors can be further investigated.

1.2 RESEARCH QUESTIONS

To bridge the gap, this research identifies a series of questions to be answered:

1. *How to detect congestion from images? How congestion can be measured from the images? How can parameters, such as crowd counts, density and flow, be related to congestion?*
2. *How to predict congestion from images? Is it possible to predict the future congestion based on current and historical congestion state? Is it advantageous to consider both spatial and temporal correlations in predictions?*
3. *How to benefit station congestion management? How to apply these models to station management? Is it feasible to implement the proposed framework in a real site?*

1.3 RESEARCH AIM AND SCOPE

In order to answer these questions, the research presented in this study aims to develop a framework for *visual congestion analysis* in both macro (congestion) and micro (pedestrian origin-destination estimation) scale, not only focusing on congestion perspective, but also exploring the possibility of other visual sensors for this research.

The overall framework of the thesis is shown in Figure 3. At the data layer, raw data (a huge amount of CCTV videos), intermediate results (frames, ground-truths) and final results (level-of-service maps and origin-destination information) are stored in the cloud server. The processing and modeling layer consists of two dimensional analysis, macro-scale and micro-scale. At a macro-scale, in order to obtain the congestion distribution and locate congestion spots, level-of-service maps of each camera view are first generated, then these camera views are combined to obtain a large map of the entire site. At a micro-scale, pedestrian origin-destination estimation is conducted to identify the bottleneck of pedestrian movement and evaluate infrastructure design. At the visualisation level, a 3D modeling environment was developed to visualise the congestion information in multi-storey or complex design stations.

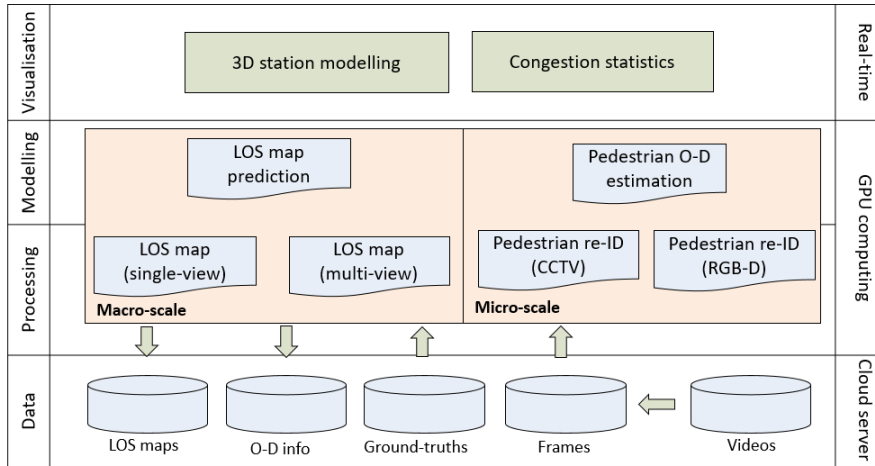


Figure 3: Research framework

1.4 MAJOR CONTRIBUTIONS

This thesis contribute towards *visual congestion analysis*. The main contributions of the study can be summarized as:

- development of spatial-temporal deep learning models for macro-scale crowd density map generation, and micro-scale pedestrian Origin-Destination (O-D) estimation based on person re-identification in the internet of cameras.
- a multi-modality framework for crowd analysis under emergency scenarios, especially low visibility situation, using RGB-D cameras.
- linking models and application by developing a web-GIS platform for real-time processing and visualization and thereby shedding more lights on the applicability of the research in crowd congestion management.

1.5 THESIS STRUCTURE

The rest of the thesis is organized as follows.

Part [i](#) explains the background and motivation of this thesis. Based on a review of current visual congestion analysis techniques and applications, Chapter [2](#) describes the concept of crowd congestion, and the up-to-date status of the visual congestion analysis in both application and technology domains, and the challenges of visual congestion analysis that hinder its further promotion.

Part [ii](#) analyzes congestion at a macro scale, Chapter [3](#) introduces a method for direct congestion map generation using a single camera in real-time. Chapter [4](#) extends the work in Chapter [3](#) into a multi-camera framework where a large camera network is used to generate a congestion map for the entire site. Based on the generated historical congestion maps, Chapter [5](#) proposes a novel congestion map prediction model using spatial-temporal deep neural network and stimulates the research from an image processing perspective.

At a micro scale in Part [iii](#), Chapter [6](#) proposes a pedestrian origin-destination estimation method based on multi-camera person re-identification for evaluating the infrastructure design. Chapter [7](#) applies RGB-D sensors for origin-destination estimation and compares the performance of RGB-D sensors in normal and emergency situations.

Part [iv](#) concludes the thesis. A general conclusion of the thesis along with the major contributions is drawn in Chapter [8](#). It also includes Chapter [9](#) of high-level discussions based on the experimental results, and projects into the future the potential trend of visual congestion analysis and technical challenges.

LITERATURE REVIEW

This chapter begins with a review of crowd congestion indicators in Section 2.1. From there on, the measurement of these congestion indicators is divided into two scales: macro scale and micro scale. At the *macro scale*, the most commonly used indicator, i.e. level of service, is discussed. Section 2.2 starts from a review of the level-of-service map generation in single camera view. Then Section 2.3 introduces techniques for fusing congestion maps from multiple camera views. Finally, Section 2.4 discusses the prediction of congestion map from an image perspective. At the *micro scale*, methods for the detection of individual pedestrians are discussed. Section 2.5 discusses the pedestrian origin-destination estimation using CCTV cameras. Then the possibility of using RGB-D cameras for this topic is reviewed in Section 2.6.

2.1 CROWD CONGESTION INDICATORS

Accurate congestion analysis relies on quantitative definition of congestion. There are four major congestion indicators in literature: level of service, degree of crowding, fundamental diagram and congestion level. Table 1 lists crowd congestion indices and parameters required to define the indices. Sources of the terms are also indicated for reference.

Level of service (LOS): Level of service is the most commonly used concept for assessing the degree of congestion in the queuing or waiting areas on the platform. Fruin and Strakosch (1987) proposed level of service as a reference to congestion. National Research Council (U.S.) (2010) defines six levels of service, ranging from LOS A (above $3.3 \text{ m}^2/\text{ped}$, least congested) to LOS F (below $0.5 \text{ m}^2/\text{ped}$, most congested) based on average pedestrian area occupancy.

Degree of Crowding (DOC): Degree of crowding or level of crowding is to evaluate the congestion inside a vehicle, e.g. train, similar to LOS, but with a different measurement location [Lam, Cheung, and Lam, 1999]. Depending on the train capacity, the degree ranges from lowest (some available seats) to highest (very

Indicators	Description	Parameters	Source
Level of service (LOS)	The concept of LOS is applied to classify the degree of congestion on the queuing or waiting areas in the platform.	Average pedestrian area occupancy	Fruin and Strakosch, 1987
Degree of crowding (DOC)	Degree of crowding or level of crowding is to evaluate the congestion in vehicles.	Train loading (percentage of train capacity) Passenger density	Lam, Cheung, and Lam, 1999
Fundamental diagram (FD)	The relation between density and flow or velocity is termed fundamental diagram to quantify the capacity of facilities.	Microscopic: individual velocity, Density	
	Network fundamental diagram (NFD) or macroscopic fundamental diagram (MFD) relates average flow variables to aggregate network flow variables.	Macroscopic: average performance (weighted flow), flow or speed, average density or number of pedestrian in the network	Seyfried, Steffen, and Lipert, 2006
Congestion level	Congestion levels are divided it into four categories from 0-3, 0 means no passenger, 3 means overcrowding	crowd density	Lo and Velastin, 2001

Table 1: Crowd congestion indicators

crowded with passengers inside the train forced to press the windows and doors).

Fundamental Diagram (FD): Fundamental diagram illustrates the relations between speed and density, which can be used to assess the design of walking infrastructure and examine the settings. Network fundamental diagram (NFD) is to evaluate the performance of the whole transportation network, which requires long time monitoring and a wide range of data [Seyfried, Steffen, and Lippert, 2006; Seyfried et al., 2005].

Congestion level: There is no official definition of this term. Lo and Velastin (2001) and Feliciani and Nishinari (2018) used congestion levels, divided into four levels, but did not describe in detail how they classified these four levels.

2.2 SINGLE VIEW CROWD CONGESTION MAP GENERATION

Image-based congestion monitoring based on CCTV images can be divided into two approaches: direct approach and indirect approach. The indirect approach tries to estimate crowd density or count the number of people based on image features, then classify these flow parameters into different congestion levels. In contrast, the direct approach classifies the image features directly to different congestion levels without calculating flow parameters.

Indirect approach

The indirect approach has been widely applied in literature since Polus, Schofer, and Ushpiz (1983) and Davies, Yin, and Velastin (1995) proposed a density metric for the level of service as the number of pedestrians per unit area, where crowd density exceeding a certain level indicates crowd congestion. Image processing techniques have been applied for the estimation of crowd density. There are several review papers on pedestrian counting and density estimation methods based on image processing techniques [Loy et al., 2013; Saleh, Suandi, and Ibrahim, 2015; Ryan et al., 2015]. Loy et al. (2013) divides counting methods into three categories: (1) counting by detection, (2) counting by clustering and (3) counting by regression, and also provides a systematic evaluation of different regression-based techniques using the same protocol. Saleh, Suandi, and Ibrahim (2015) reviews the crowd density estimation and counting methods for visual surveillance and divides them into two approaches: direct approach (object-detection

based) and indirect approach (pixel-based, texture-based and corner points based). Ryan et al. (2015) evaluates holistic, local, and histogram-based methods and compares various image features and regression methods using several public datasets.

Direct approach

While a large number of studies adopt the indirect approach based on flow parameters, little research has focused on the direct measurement of congestion. Lo and Velastin (2001) proposes a multi-layer feed forward neural network for classifying the levels of congestion from CCTV cameras. But this research regards the whole camera view as one congestion level, which provides a low-resolution congestion map and can not assist the station managers in locating the congestion spots accurately. Most of existing methods for congestion measurement in images are indirect (which has disadvantages) and the few ones that are direct assign one congestion level to the whole image. Therefore a method for measuring different congestion levels directly in one or more image does not exist.

Image classification

Image classification is used to directly classify the images into different levels of service. The general procedure of *Image classification* contains two steps. The first step is to retrieve features from images, and the second step is to construct a classification model using the extracted features as inputs. Deep neural networks (DNNs) are a special set of machine learning methods that have achieved remarkable success in various image processing tasks. Convolutional Neural Networks (CNNs) [Schmidhuber, 2015] show the state-of-the-art performance in ImageNet challenge and outperform humans in image classification task. Since the training of a CNN requires large training datasets and computing resources, many pre-trained models have been made publicly available, including VGG-VD [Simonyan and Zisserman, 2014], AlexNet [Krizhevsky, Sutskever, and Hinton, 2012], ResNet18, ResNet50 [He et al., 2016], DenseNet161 [Huang et al., 2017]. Recurrent neural networks (RNNs) have been shown to be successful in time sequence prediction tasks [Sutskever, Vinyals, and Le, 2014]. Long short-term memory (LSTM) is a special RNN that can learn long-term temporal dependencies, for example the temporal relations between image frames. As for the classification model, the support vector machine (SVM)

[Cortes and Vapnik, 1995] can capture main features that characterize the algorithm (maximal margin). It is particularly suitable for high dimensional classification problems with a limited number of training samples.

2.3 MULTI-VIEW CROWD CONGESTION MAP GENERATION

Multi-camera correspondence

When dealing with multiple camera views, the correspondence between multiple camera views needs to be established. Conventional methods for establishing camera correspondence can be classified into two categories: homographic-based and calibration-based methods [Tang et al., 2015].

Most homographic-based methods, calculate plane homographies by matching discriminative features across images, e.g., SIFT features [Li, Huang, and Liu, 2012], foreground polygons [Mousse, Motamed, and Ezin, 2017], or corresponding points, e.g., people heads [Ma, Zeng, and Ling, 2012], feet [Maddalena, Petrosino, and Russo, 2014], and then establish the correspondence among multiple cameras using homography transformation.

The objective of calibration-based methods [Liu, Collins, and Liu, 2011; Gardziński et al., 2015] is to determine the perspective model of a camera, which includes 1) the extrinsic parameters, i.e., the position and orientation of the camera in the real-world coordinate system; and 2) the intrinsic parameters, i.e., the optical center and focal length of the camera. Since the camera parameters are not necessary in this study, the homographic-based method is more suitable for our work.

Ensemble learning

Ensemble learning, a category of machine learning techniques, is a process of solving a particular computational intelligence problem by strategically combining multiple models, for example classifiers or experts, which has shown favorable results compared to those of single-source system for a broad range of scenarios, e.g., transportation, financial, medical, social area [Opitz and Maclin, 1999; Polikar, 2006; Rokach, 2010; Azadbakht, Fraser, and Khoshelham, 2018]. Multi-view fusion is one of the important applications of ensemble learning, which receives data obtained from various sources that may provide complementary information [Atrey et

al., 2010; Snoek, Worring, and Smeulders, 2005; Hall and Llinas, 1997].

Two key components of ensemble learning are first, to build diverse classifiers using the popular bagging [Breiman, 1996], boosting, AdaBoost [Freund, Schapire, and Abe, 1999], stacked generalization [Wolpert, 1992], or mixture of experts algorithms [Jacobs et al., 1991]. In order to achieve higher performance, the classifiers should be as diverse as possible. If similar classifiers are generated, ensemble learning would not work very well for the whole process. The second component is to merge the classifier results based on a combination rule. Several combination rules have been used, including algebraic combination of outputs, voting based techniques and borda count.

From the perspective of levels of fusion, there are two levels of fusion: early fusion and late fusion [Snoek, Worring, and Smeulders, 2005]. Early fusion is to fuse the information at feature level. After analysis of the various unimodal views, the extracted features are combined into a single representation, and then fed to learn the concepts. The disadvantage of the early fusion approach is the difficulty of combining features into a common representation. Late fusion is to fuse multiple modalities in the semantic space or at decision level, which learn semantic concepts directly from unimodal features. These learning results and scores are combined afterwards to yield a final detection score. Late fusion focuses on the individual strength of modalities. As our previous work [Li et al., 2018b; Li et al., 2019a] has shown that single view congestion classification achieves a high accuracy, late fusion is more suitable for the combination of congestion classification results from multiple views. In this thesis, we compare and evaluate different combination rules.

2.4 CONGESTION PREDICTION

Spatio-temporal prediction

The patterns in traffic prediction can be divided into three categories [Zhang, Zheng, and Qi, 2017]: 1) Spatial relations: The traffic performance, for example traffic flow, is affected by the nearby regions, more than the remote areas. 2) Temporal relations: The traffic performance is also related to both near future and far future. Traffic flow changes gradually in continuous time and also

shows periodic patterns. The morning and afternoon peak hours may have similar patterns on weekdays. 3) External features: Some external factors, such as weather conditions and road works will affect the traffic performance.

The general procedures of spatio-temporal traffic prediction using deep learning are composed of two steps: to first extract the features, then relate the features to the traffic parameters we intend to predict. Deep learning methods have been applied to extract different features depending on the characteristics of the neural networks. Typically convolutional neural network (CNN) is used to extract spatial features and recurrent neural network (RNN) is used to extract temporal features. CNN models use convolutional operators to learn image features and can preserve the spatial relations between pixels, achieving remarkable performance on image recognition [Krizhevsky, Sutskever, and Hinton, 2012]. Different structures of CNN have been applied, for example standard CNN [Wu and Tan, 2016; Ke et al., 2017; Zhang, Zheng, and Qi, 2017], graph CNN [Yu, Yin, and Zhu, 2017]. RNN models are good at handling temporal tasks because they recurrently calculate each element of time series data and memorize the captured information. Gated residual unit (GRU) [Li et al., 2018a] and long short-term memory (LSTM) [Yu, Yin, and Zhu, 2017; Wu and Tan, 2016] are two special categories of RNN, which both can learn long-term temporal dependencies. While GRU has two gates (update and reset), LSTM has three gates (forget, input and output) and thus more parameters than GRU resulting in better performance on big volume of data. The combination of CNNs and RNNs is emerging as a new popular approach as the hybrid model inheriting the advantages of CNNs and RNNs can capture both the spatial and temporal traffic correlations at the same time [Yu et al., 2017a]. Convolutional long short-term memory (ConvLSTM) [Xingjian et al., 2015] is a promising structure that can learn spatial and temporal dependencies simultaneously.

To relate the extracted features to the traffic parameter we intend to predict, some works concatenate the spatial and temporal features first and then build a regression model to predict, for example linear regression [Wu and Tan, 2016], logistic regression [Lv et al., 2015], support vector machine [Li et al., 2018b] and random forest [Ke et al., 2017]. Other works pass the concatenated features to a fully-connected layer and generate the final prediction result [Yu et al., 2017a].

Image sequence prediction

Image sequence prediction can be divided into two tasks: high-level semantic prediction and pixel-level prediction [Villegas et al., 2017]. High level semantic prediction refer to action, event and motion. Forecasting semantics describes what will happen and provides basis for decision-making. The training for semantic prediction tasks requires a large amount of training data, which involves a tremendous manual labeling work. In contrast, pixel-level prediction (also called frame prediction) provides a dense and direct description of the world. Spatio-temporal correlations in videos provide a self-supervision for frame prediction, which enables purely unsupervised training of a model by observing raw video frames.

Deep neural networks, which are similar to the certain areas of the brain and can be trained to decompose signals, preserve information, and perform classification, have dominated the image sequence prediction task. CNNs can find spatial patterns and decompose the inputs into simpler representations, so that they are very useful for the identification of objects and movements [Batchuluun et al., 2018]. RNNs can remember sequences and predict the next movement [Alahi et al., 2016]. A mixture of them allows a computer to recognize, remember and make predictions according to the current and previous inputs [Balderas, Ponce, and Molina, 2019].

2.5 PEDESTRIAN ORIGIN-DESTINATION ESTIMATION

Pedestrian O-D estimation is mainly based on trajectories. *Pedestrian tracking* for trajectory estimation in single camera views has been a topic of intensive research in computer vision. Boltes and Seyfried (2013) discussed the methods of collecting pedestrian trajectories under laboratory settings where overhead cameras perpendicular to the floor allow a view without occlusion. Zaki and Sayed (2017) localized pedestrians of small groups in the real environment for analyzing pedestrian walking behaviour. Due to the insufficiency of single view tracking in large-scale pedestrian facilities, pedestrian tracking across disjoint cameras developed to an independent topic *pedestrian re-identification* (re-ID). Different from *pedestrian tracking*, the key challenge in pedestrian re-ID is to match two images of a person without spatial and temporal cor-

relations but with intensive appearance changes, such as lighting, pose and viewpoint.

Generally, pedestrian re-ID is composed of two main procedures: first to generate unique person descriptors, second to compare the descriptors of a *query* with one or two images in the *gallery* to infer either a match or a non-match. Descriptors are descriptions of the visual features of the pedestrians in images, such as shape, color, texture, motion information, which are usually in the form of feature vectors. Depending on the categories of descriptors, the pedestrian re-ID system can be divided into hand-crafted systems and deep learning based systems [Zheng, Yang, and Hauptmann, 2016]. In hand-crafted systems, color information is most frequently used, represented by color histograms, and texture features can be extracted by filter templates. As hand-crafted features are variant in different camera environments, a good distance metric is quite important for matching in hand-crafted systems. The purpose of distance metric learning is to keep all the feature vectors of the same class closer while vectors of different class fall apart. The most popular formulation is based on Mahalanobis distance which calculates the Euclidean distance between feature vectors for generating a pairwise distance matrix between the *query* and each *gallery* images, and then find the pair with the minimum distance [Yang and Jin, 2006].

As deep neural networks (DNN) outperform humans in image recognition tasks, they have been applied to extract accurate and robust person descriptors in pedestrian re-ID. Two types of DNNs have been employed for person re-ID: 1) image classification models; 2) siamese models based on pair or triplet comparisons. Under the condition that how many identities (class) are available, a pedestrian re-ID model based on image classification determines the individual identity that the *query* image belongs to. A limitation of this method is that it requires a fairly large number of training images to avoid the over-fitting problem that can affect the performance of the method. The siamese networks have been proposed to solve the problem of insufficiency of training samples. Siamese network is a type of neural network structure that contains two or more identical sub-networks. Pair-wise models take two images as input and two sub-networks are connected with pairwise loss function while triplet models take three images as input and three sub-networks are connected by triplet loss function [Hermans, Beyer, and Leibe, 2017].

To overcome the challenges of pedestrian re-ID, including variations in human pose and illumination, occlusion and background clutter, learning robust features is quite important. Three types of features have been discussed in literature: global features, local features and attributes. In general, global features emphasize more on the color and texture features for all the regions. ResNet-50 [He et al., 2016], a residual learning framework has been proposed to learn global features and more importantly to ease the training of networks that are substantially deeper than those used previously and gain accuracy from considerably increased depth. Yu et al. (2017b) argue that both high and mid-level global features are relevant for cross-domain instance matching and proposed a unified framework based on ResNet-50 to fuse the high-level features and middle-level features in the early layers for classification.

Recent studies focus on improving global representation by leveraging local features extracted from human body parts [Bai et al., 2017; Sun et al., 2018; Suh et al., 2018; Kalayeh et al., 2018]. Bai et al. (2017) regards the pedestrian as a sequence of body parts from head to foot, and applies Long Short-Term Memory (LSTM) in an end-to-end fashion to take into account the contextual information between body parts, enhancing the discriminative capacity of local feature which aligns better to full person. Sun et al. (2018) proposed a network named Part-based Convolutional Baseline (PCB) which conducts uniform partition on the conv-layer for learning local features. Suh et al. (2018) handles the body part misalignment problem, that is, body parts are misaligned across human detections due to pose/viewpoint change and unreliable detection and shows that the part-aligned representation leads to a robust image matching similarity. Different from previous studies based on bounding box detection, Kalayeh et al. (2018) adopts human semantic parsing to segment accurate body regions, namely foreground, head, upper-body, lower-body and shoes, and then use them to exploit local cues for person re-identification.

Other research works such as Si et al. (2018) and Li, Zhu, and Gong (2018) develop attention-based models to automatically select more discriminative body parts. Si et al. (2018) proposed an end-to-end trainable framework to learn context-aware feature sequences and perform attentive sequence comparison simultaneously with dual attention mechanism. Li, Zhu, and Gong (2018) presented a novel Harmonious Attention Convolutional Neural

Network (HA-CNN) for joint learning of attention selection and feature representations in an end-to-end fashion.

Attributes as middle-level representations of pedestrians have been popular recently. Chang, Hospedales, and Xiang (2018) proposed a Multi-Level Factorisation Net (MLFN), a CNN architecture that factorises the visual appearance of a person into latent discriminative factors at multiple semantic levels without manual annotations. The factors, such as gender, carrying, clothing, texture, color, etc., are computed at different levels of the network corresponding to latent attributes of different semantic levels.

Although many methods have been proposed to tackle the problem of pedestrian re-ID, whether they can handle diverse complicated scenes with diverse densities and be applied for crowd congestion management remains unsolved. To fill this gap, in this study we evaluated these state-of-the-art methods in diverse congestion levels and improved the performances of the models by data augmentation techniques.

2.6 CROWD ANALYSIS USING RGB-D CAMERAS

Due to the emergence of cost-effective 3D sensors, RGB-D data have been applied to crowd related analysis, e.g., face recognition, human recognition and actions identification [Firman, 2016], which has been proved to be more insightful and privacy friendly than 2D data. For example, 3D depth cameras were installed in Brisbane central rail station for people detection and tracking and reduce passenger dwelling time in the platforms [Virgona, Kirchner, and Alempijevic, 2015].

RGB-D sensors

Most research apply kinect sensor to capture depth data, but limited to the capture distance and viewing angle, the moving distance of people is constrained. With the fast development of RGB-D based pedestrian detection, depth data has received significant attention. Table 2 compares state-of-art 3D consumer cameras. Zed sensor from Sterolabs is a good choice, which can capture 110° wide-angle video and achieve up to 20m depth perception both indoors and outdoors.

RGB-D based pedestrian recognition

People recognition related tasks include pedestrian detection, track-

ing and re-identification. In computer vision, people detection is to recognize the biometrics characteristics of people and locate the people in the frames. On the basis of pedestrian detection, pedestrian tracking is matching and tracking the same person in continuous space and time, while people re-identification is identifying the same people in either different space or different time.

All these datasets were collected in the campus, e.g., canteen, hallway, lab, lecture hall. Two datasets, RGB-D people [Spinello and Arras, 2011] and OPR [Lejbølle et al., 2017], shot in canteen are under uncontrolled environment. Increasing and decreasing flows during lunch time were recorded. Multiple persons appeared in one frame and various flow densities were recorded. While other eight datasets were created under controlled environment. People in isolation were instructed to conduct normal walk, standing still or stretching their arms. As for the number of participants, TUM-GAID dataset [Hofmann et al., 2014] recorded 305 persons. RGB-D people dataset recorded 1648 human instances, but how many people are the same identities are not clear. Other eight datasets recorded less than 100 identities.

Pedestrian recognition research can be divided into long-term and short-term. In long-term people recognition [Barbosa et al., 2012; Munaro et al., 2014a; Hofmann et al., 2014; Pala et al., 2016; Liciotti et al., 2016], people change their appearance when walking through the cameras, e.g. changing clothes, carrying bags, changing shoes. For example, TVPR dataset [Liciotti et al., 2016] shot eight days so that the monitored persons change their clothes in these days. Short-term research [Spinello and Arras, 2011; Munaro et al., 2014b; Kastaniotis et al., 2015; Haque, Alahi, and Fei-Fei, 2016; Lejbølle et al., 2017] are often appearance-based because the looks of people are unchanged.

We note that all the datasets apply single camera to record the movements or shoot in one location. Although RGB-D people dataset [Spinello and Arras, 2011] binded three cameras together in one location to achieve a e view of the canteen hall, it is still single camera view. As for the camera angle, most datasets are shot from front view, back view or side view, in order to have a clear view of the whole human body. Later top-view configuration appears to avoid people occlusion and protect the privacy of humans, e.g., DPI-T [Haque, Alahi, and Fei-Fei, 2016], TVPR [Liciotti et al., 2016] and OPR [Lejbølle et al., 2017] datasets.

Pedestrian behaviour analysis using RGB-D data

The activities studied can be divided into three categories: daily activities, health-related activities and mutual activities. Daily activities are those routine activities in the office or in the lab, including drinking water, eating, reading, etc. Most datasets include this category. Monitoring health-care related activities monitoring is very important for medical care, and therefore has aroused great interests, including sneezing, staggering, falling down, etc. Mutual activities, including punching, kicking, hugging, etc, are studied in three datasets [Wang et al., 2014c; Liu et al., 2017; Shahroudy et al., 2016]. The number of activities in these datasets ranges from 8 to 60. We note that all of these datasets are collected under controlled environment, which means that every participants are instructed to conduct each activity. As deep learning has shown remarkable performance in many tasks and been applied to behaviour analysis, large-scale datasets are needed for training. The participants conduct each action several times so that several videos of each action are collected. The videos labeled for activity recognition ranges from 60 to 56880 clips.

Due to view variations, which are common in real conditions, visual appearances of humans are different in different views. Multi-view or cross-view datasets are desirable as recognizing actions from the videos which are unseen in the training videos is necessary [Wang et al., 2014a]. Act 4² dataset [Cheng et al., 2012] is the first cross-view RGB-D dataset. NTU RGB dataset [Shahroudy et al., 2016] set 80 views for the experiment to test the multi-view performance. All the cameras are set aside the objects so that the whole human actions can be seen without occlusion.

Multi-modality is another trend in RGBD datasets. Microsoft kinect v1 provide RGB data and depth data. With Microsoft kinect SDK and Open NI SDK, the human skeleton data can be extracted. As Microsoft kinect v2 sensor also provides infrared data, two datasets [Liu et al., 2017; Shahroudy et al., 2016] using this sensor add this modality, thus four modalities in total are available, including RGB, depth, skeleton and IR data.

From the perspective of transportation application, two surveys summarized the studies on people behaviour recognition in transit scenes [Candamo et al., 2010] and road intersections [Shirazi and Morris, 2017]. However, current studies are all RGB based. None of these analysis are RGB-D based. For example, Candamo et al. (2010) divides the human behaviours in transit scenes into single person behaviours (e.g., loitering), multiple-person inter-

actions (e.g., fighting and person attacks), person-vehicle interactions (e.g., vehicle vandalism), and person-facility/location interactions (e.g., object left behind and trespassing). Shirazi and Morris (2017) analyzes three participants at intersections: vehicles, drivers and pedestrians and assess their behaviours and safety. Pedestrian related behaviour assessment include motion prediction, waiting time, route/crossing choice, and crossing/walking speed. Safety analysis, include gap analysis, conflict analysis between vehicle and pedestrian, risk and accident possibility at intersections. However, as listed in Table 4, current RGB-D datasets cannot support the above behaviour analysis in transport scenes, as all the actions in these datasets are not designed for transport applications. Thus, a new RGB-D dataset for people action analysis is desirable.

Picture				
Company	Microsoft	ASUS	Intel	Stereolabs
Brand	Kinect 2.0	XtionPro2	RealSense D415/435	ZED
Depth range	0.5-4.5 m	0.8 to 3.5 m	0.3 to 10 m	1.5 to 20 m
3D Resolution	512 x 424	640 x 480	1280 x 720	2208 x 1242
RGB Resolution	1920 x 1080	1280 x 1024	1920 x 1080	2208 x 1242
Frame Rate	30 fps	30 fps	90 fps	15 fps
FOV	70 H, 60 V	58 H, 45 V	69.4 x 42.5 x 77(415) 86 x 57 x 94 (435)	96 H, 54 V, 110D
Developer kit	Microsoft SDK 2.0		RealSense 2.0 SDK	ZED SDK 2.3
Price	used: \$60-180	used: \$129-155	\$149-179	\$449

Table 2: List of 3D consumer cameras

Datasets	Place	Scene	Density	IDs	Cams	Angle	Device	Year
RGB-D people[Spinello and Arras, 2011]	canteen	uncontrolled	multiple		3 in 1	horizontal	kinect	2011
PAVIS[Barbosa et al., 2012]	lab	controlled	isolation	79	1	horizontal	kinect	2012
IAS-Lab[Munaro et al., 2014a]	lab	controlled	isolation	11	1	horizontal	kinect	2014
BIWI[Munaro et al., 2014b]	lab	controlled	isolation	50	1	horizontal	kinect	2014
TUM-GAID[Hofmann et al., 2014]	corridor	controlled	isolation	305	1	horizontal	kinect	2014
UPCV gait [Kastaniotis et al., 2015]	lab	controlled	isolation	30	1	horizontal	kinect	2015
KinectREID [Pala et al., 2016]	classroom	controlled	isolation	71	1	horizontal	kinect	2016
DPI-T [Haque, Alahi, and Fei-Fei, 2016]	hallway	controlled	isolation	12	1	overhead	kinect	2016
TVPR[Liciotti et al., 2016]	hallway	controlled	isolation	100	1	overhead	ASUS	2016
OPR[Lejbølle et al., 2017]	canteen	uncontrolled	multiple	64	1	overhead	ZED	2017

Table 3: RGBD datasets for pedestrian recognition

Datasets	Acts	Actno	Density	IDs	Videos	Views	Modality	Year
MSR-Action3D[Li, Zhang, and Liu, 2010]	daily	20	isolation	10	567	1	D+skeleton	2010
CAD-60[Sung et al., 2011]	daily	12	isolation	4	60	1	RGB+D+skeleton	2011
RGBD-HuDaAct[Ni, Wang, and Moulin, 2011]	health	13	isolation	30	1189	1	RGB+D	2011
MSRDailyAct3D[Wang et al., 2012]	daily	16	isolation	10	320	1	RGB+D+Skeleton	2012
Act4 ² [Cheng et al., 2012]	daily	14	isolation	24	6844	4	RGB+D	2012
UTKinectAction [Xia, Chen, and Aggarwal, 2012]	daily	10	isolation	10	200	1	RGB+D+Skeleton	2012
CAD-120[Koppula, Gupta, and Saxena, 2013]	daily	20	isolation	4	120	1	RGB+D+Skeleton	2013
3D Action Pairs[Oreifej and Liu, 2013]	daily	12	isolation	10	360	1	RGB+D+Skeleton	2013
DMLSmartAction[Amiri et al., 2013]	daily	12	isolation	16	932	3	RGB+D	2013
MHAD[Ofli et al., 2013]	daily	11	isolation	12	660	4	RGB+D+Skeleton	2013
Multiview3Devent[Wei et al., 2013]	daily	8	isolation	8	3815	3	RGB+D+Skeleton	2013
OnlineRGBDAction[Yu, Liu, and Yuan, 2014]	daily	7	isolation	24	336	1	RGB+D+Skeleton	2014
NorthwesternUCLA[Wang et al., 2014a]	daily	10	isolation	10	1475	3	RGB+D+Skeleton	2014
UWA3Dmultiview[Rahmani et al., 2014]	daily	30	isolation	10	900	1	RGB+D+Skeleton	2014
office activity[Wang et al., 2014c]	daily	20	1-2 persons	10	1180	3	RGB+D	2015
TJU[Liu et al., 2015]	daily	22	isolation	21	1936	1	RGB+D+Skeleton	2015
UTD-MHAD[Chen, Jafari, and Kehtarnavaz, 2015]	daily	27	isolation	8	861	1	RGB+D+Skeleton	2015
UWA3Dmultiview2[Rahmani et al., 2016]	daily	30	isolation	10	1075	5	RGB+D+Skeleton	2015
NTU RGB[Shahroudy et al., 2016]	40daily							
	9health	60	1-2 persons	40	56880	80	RGB+D+IR+Skeleton	2016
	11mutual							
PKU-MMD[Liu et al., 2017]	41daily	51	1-2 persons	66	1076	3	RGB+D+IR+Skeleton	2017
	10mutual							

Table 4: Datasets for behaviour analysis

Part II

MACRO-SCALE CROWD CONGESTION MAP GENERATION

This part focuses on crowd congestion measurement at macro-scale, i.e. estimating the density rather than the number of individuals. At the macroscopic level, Chapter 3 introduces a direct congestion map generation method in real-time. Chapter 4 extends this method to large camera networks, considering the large camera networks and generating congestion map in the whole space. Based on the produced historical congestion maps, Chapter 5 proposes a novel congestion map prediction model using spatial-temporal deep neural networks and stimulates the research from an image processing perspective.

SINGLE VIEW LEVEL OF SERVICE MAP GENERATION

This chapter is based on two articles: one journal article published in *Journal of Intelligent Transport Systems* [Li et al., 2019a] and a peer-reviewed conference paper presented at *Transportation Research Board Annual Meeting* [Li et al., 2018b]. As the first author, I proposed the research idea, came up with the hypothesis, designed and implemented the experiment, and conducted the data analysis, discussions and conclusions. My supervisors Prof. Majid Sarvi, Dr. Kourosh Khoshelham and Dr. Milad Haghani were actively involved in supervising me, discussing the research idea and design, and also proofreading the paper.

3.1 INTRODUCTION

A pedestrian congestion map is a graphical, real-time or near-realtime representation of pedestrian flow for particular areas of interest. The map can help station operators to locate congested spots and manager the crowds. Pedestrian level of service (LOS) is the most widely accepted standard for measuring congestion. However, as reviewed in Chapter 2.2, existing methods for measuring LOS based on crowd density estimation from images have the disadvantages that, crowd density cannot be estimated accurately. In addition, the calculation of flow parameters involves a complex process, and consequently these parameters are not indicative of congestion in real-time.

In this chapter we propose a new approach to automatic congestion monitoring using CCTV images. Our approach is based on convolutional neural networks (CNNs) to automatically learn visual features of the crowd, the support vector machine (SVM) to categorize LOS and long short-term memory (LSTM) to refine the results based on temporal relations between frames. The contributions of the present research are the following:

1. A direct method for congestion classification in CCTV images based on CNN and SVM.

2. A method for refining LOS maps by exploiting the temporal correlation between successive images using the LSTM network.
3. A web-based LOS map visualization enabling the analysis of congestion distribution and its variation in real-time.

This chapter is organized with the following sections. Section 3.2 depicts the conceptual framework and Section 3.3 gives details of the experiment implementations and results, followed by section 3.4 of conclusion and discussions.

3.2 OVERVIEW OF THE APPROACH

Figure 4 shows the five primary steps of our approach. First we define the region of interest (ROI) of the platform and create grids corresponding to the unit area. Then we feed the grids to CNN for image features extraction, and SVM is adopted for classifying the features into different levels of service. Furthermore results refinement is conducted by considering the correlations between successive frames. At last LOS maps are visualized in a web-based platform.

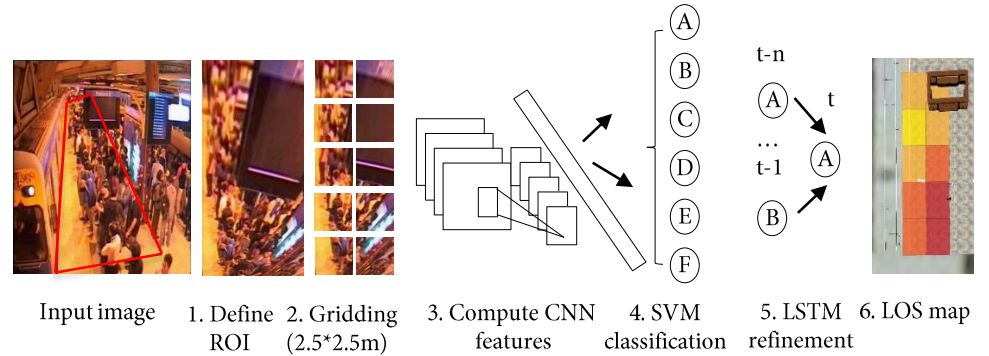


Figure 4: Overview of the approach

3.2.1 Defining the region of interest

We define the level of service for a rectangular area on the ground. When projected to the image, this rectangular area becomes a trapezoid due to the perspective geometry of the image. To be able to apply the feature extraction and classification, we employ a geometric transformation to create a rectangular image from the trapezoid region of interest, as shown in Figure 5.

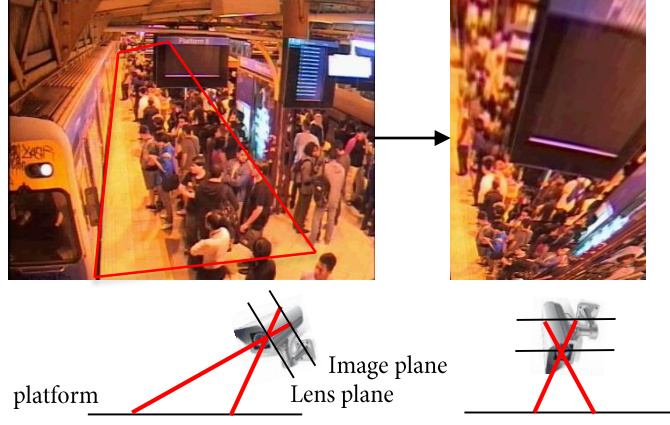


Figure 5: Geometric transformation of the region of interest into a rectangular image

The region of interest is initially defined in a 3D model of the station we have created for the purpose of this research. Correction of perspective distortion is done by collecting 4 corner points of the platform in the 3D model and their corresponding coordinates in the image space. A homography matrix is then constructed to map the selection area to a rectangular shape. The resulting image includes the platform as the detection area and objects outside the platform are filtered out as their textures may cause confusion and reduce the detection accuracy.

3.2.2 Creating grids

For practical applications, it is important that a unit area for the analysis of the level of congestion is defined. Accordingly, the region of interest is subdivided into a grid of cells corresponding to this unit area. Typically, cells of an area of approximately 10 m^2 are used in crowd dynamics (Asano et al., 2007; Guo, Huang, and Wong, 2011). Lo and Velastin (2001) adopts the whole camera view to calculate LOS. However, as shown in Figure 5, one camera view can cover a large area, more than $20\text{m} \times 20\text{m}$. This results in a coarse resolution for the congestion analysis, as different congestion levels can be present in one camera view.

Based on practical requirements for congestion monitoring and the width of each platform in our experiment, which is 5 meters, we choose cells of $2.5\text{m} \times 2.5\text{m}$ as the unit analysis area, to which one level of service will be assigned. On the marked grid of 2.5m squares, a density lower than 1 person per grid cell indicates free

flow of the crowd, LOS A, as illustrated in Table 5. As the number of people per grid cell increases, movement starts to become constrained. A density of 13 people or more per grid cell indicates, LOS F, which is the most congested situation.

Table 5: Level of service: the most widely accepted congestion index (Source: Highway capacity manual (National Research Council (U.S.) (2010)))

LOS	Average area occupancy (m^2/ped)	No. of people (area of $2.5\text{m} \times 2.5\text{m}$)	Description
A	3.3 or more	0 - 1	Free flow, convenient passing, conflicts avoidable
B	2.3 - 3.3	2	Minor conflicts, passing and speed restriction
C	1.4 - 2.3	3 - 4	Crowded but fluid movement, passing restricted, crossing and reverse flow difficult
D	0.9 - 1.4	5 - 6	Significant conflicts, passing and speed restriction, intermittent shuffling
E	0.5 - 0.9	7 -12	Shuffling walk; reverse, passing and crossing very difficult, intermittent stopping
F	0.5 or less	13 or more	Critical density, flow sporadic, frequent stops, contact with others.

3.2.3 Feature extraction and classification

Each sub-image corresponding to a grid cell is fed to a CNN for feature extraction and classification. In this paper we apply a pre-trained CNN (VGG-VD model) with 16 weight layers (Simonyan and Zisserman, 2014). This model has been trained on a subset of the ImageNet database with more than a million images and can classify images into 1000 object categories. The CNN structure as

shown in Figure 6 consists of 16 layers with learnable weights: 13 convolutional layers, and 3 fully connected layers. The last layer of the model (softmax) provides the classification results. In this paper we replace the softmax layer with a SVM classifier, because SVM is more suitable for high dimensional classification problems with a limited number of training samples. This research builds the SVR model by taking image vectors as inputs and level of service as outputs. First training images are partitioned into 10 parts to train the classification model and the best fit SVM model is decided after 10-fold validation on the training data. Then we perform classification on test data based on the trained model.

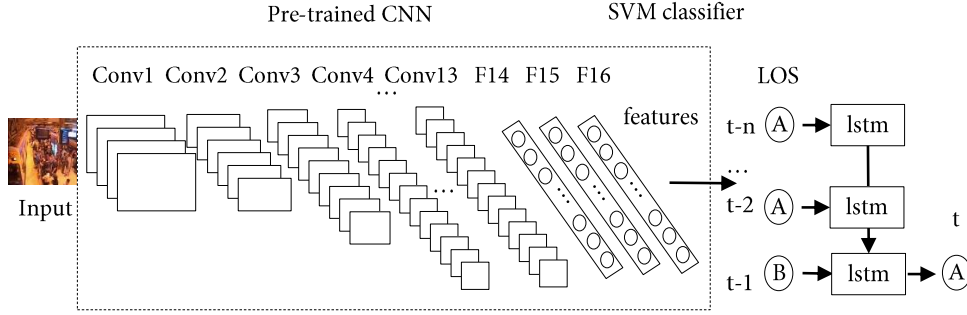


Figure 6: Structure of deep learning network for LOS classification

3.2.4 Refining level of service labels

As the CNN-based classification considers each frame independently, the classification results can fluctuate dramatically between adjacent frames. Inspired by Tan, Zhang, and Wang (2011) and Wang et al. (2014b), which have pointed out that the temporal correlations of pedestrian counting results in improved estimates for the detection of pedestrians, we adopt the LSTM method to refine the time-series classification results.

As shown in Figure 6, the LSTM in this structure is a one layer sequential model. We used 50 LSTM nodes in the layer to which we gave input of shape (2,1), which is two previous frames given to the network. Generally current values are most affected by most recent data. After comparing the results of different number of input frames, one input with two previous frames generates the best refinement result (The frame interval is 1 second). The last layer is a dense layer where the loss is mean absolute error with adaptive moment estimation (Adam) as an optimizer. We train this model for 200 epochs.

3.2.5 Web-based visualization of level of service maps

In order to visualize the congestion levels and better locate the congested area, a heat map is frequently used to represent the congestion distribution. But the problem with heat map is that managers can not identify the actual location of the congestion. To alleviate this issue, heat maps are usually overlaid on a spatial map such as a floor plan. In this paper, we propose to overlay the congestion levels on a 3 dimensional model. As many train stations are multi-storey buildings, visualization of the congestion on a 3 dimensional model allows for more convenient monitoring of the congestion in the whole station.

Figure 14 shows the web-based platform we developed. We use a Cesium platform to provide browser-based 3D visualization and congestion analysis. The station manager can choose the platform they intend to monitor by clicking the platform selection button. Then the video player window plays current CCTV videos and LOS maps overlaid on a 3 dimensional model is synchronized to show the LOS maps in real-time. Time slider is provided to display the current time period. Map navigation tools for interactive maps are also provided. With the pan and zoom tools, station managers can browser the level of service maps in multi-storey stations.

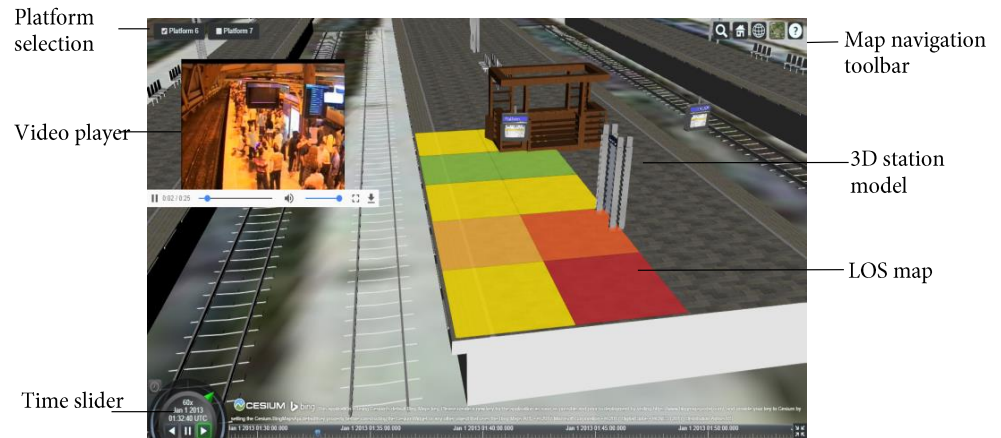


Figure 7: A Cesium-based interface is developed to visualize interactive LOS maps. The 3 dimensional model of Flinders Street Station is presented to better locate the congestion positions.

3.3 EXPERIMENTAL EVALUATION

3.3.1 *Data sets*

The CCTV data captured on platform 6 of Flinders Street Station in Melbourne, Australia was used for the evaluation of the method. The video period is from 01:28 am to 01:48 am during the 2013 New Year Eve, and shows the gradual increase and decrease of pedestrian flow on the platform, providing a perfect coverage of LOS from A to F.

In order to collect ground truth data, we developed an interface for manually labeling the position of people in each grid and invited 3 reviewers to label the same dataset. This allowed us to obtain average crowd counts so as to reduce the error caused by human recognition. Then we converted crowd numbers into ground-truth LOS according to Table 5.

3.3.2 *LOS classification accuracy*

The classification methodology is implemented using the neural network toolbox and LibSVM packages in MATLAB. A 10-fold cross validation schema was employed to evaluate the classification accuracy. The classification accuracy is shown in Table 6. The total accuracy for all samples is 81.9%. Furthermore, the samples of LOS F shows the best performance compared with other levels of service, which indicates that our approach produces better results in more crowded scenes.

The results on part of the test dataset is shown in Figure 8. The classification levels with LSTM refinement (green line) are closer to the ground-truth counting (blue line) than those without LSTM refinement (red line). Considering the temporal correlations between frames, there are fewer fluctuations of congestion levels around the ground-truth.

3.3.3 *Computational efficiency*

The algorithm were run on a 2.6-GHz Intel Core i7 processor with 16GB of memory. The average processing time are shown in Table 7. The average total processing time for each frame is 0.40 seconds, which allows near real-time classification of congestion levels in CCTV images.

Table 6: Number of samples and accuracy for each level

	LOS A	LOS B	LOS C	LOS D	LOS E	LOS F
Example						
No of samples	293	509	2165	574	1243	256
Accuracy	81.5%	84.1%	85.7%	66.7%	78.8%	95.3%
Confusion matrix						
Predicted label						
LOS A	239	8	22	0	0	0
LOS B	38	428	140	19	11	0
LOS C	16	69	1856	120	57	0
LOS D	0	3	128	383	123	1
LOS E	0	0	15	50	979	11
LOS F	0	1	4	2	73	244
Total accuracy	81.9%					

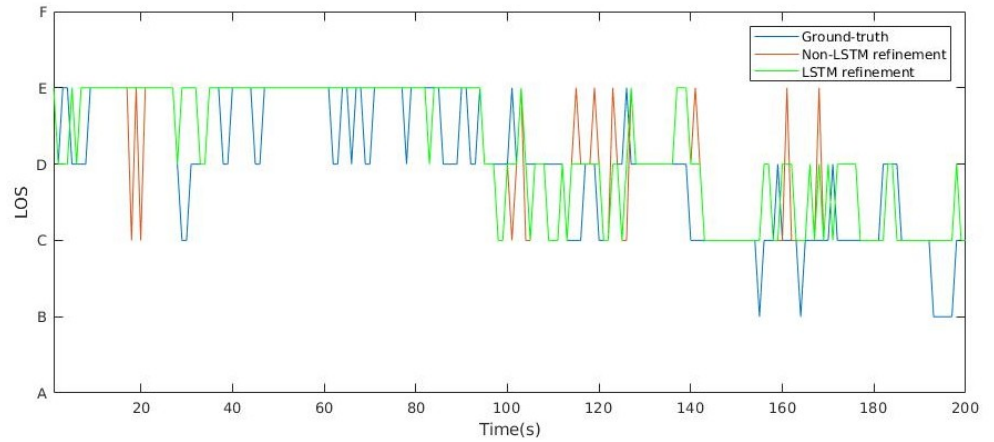


Figure 8: LSTM refinement method (green line) which considers temporal relations is more close to the ground-truth (blue line) than non-temporal method (red line), which is more fluctuate around ground-truth.

3.3.4 Temporal analysis of LOS

Figure 9 shows the level of service maps generated by our method in 5 minute interval between 01:28 am and 01:48 am. The train arrives at 01:35 am and boards the passengers. Consequently the platform becomes empty. This can be clearly observed in the third LOS map, where grid cells in green indicate level of service A,

Table 7: Average processing time for each frame

Procedures	Time
Define ROI	0.88
Gridding	1.36
Feature extraction	370.87
Classification	1.88
Optimization	15.08
Visualization	0.15
Total time (ms)	390

least congested, so the third graph becomes all green, which is the lowest level of service and soon the level of service increases as new passengers enter the station.

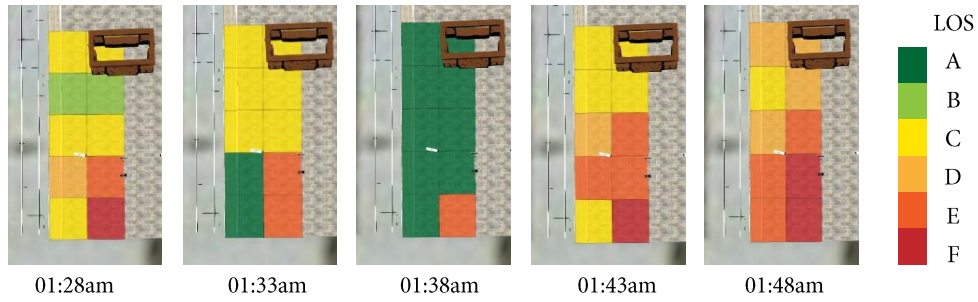


Figure 9: level of service map update in 5-minute intervals

From Figure 10, the variation of congestion in each grids is presented, which could be valuable for the analysis of occupancy in different zones. It can be observed that pedestrians generally tend to stand in the inner side of the platform as the average level of service in the inner side (grid cells 02, 04, 06, 08 and 10) shows higher congestion levels than that in the outside grids (grid cells 01, 03, 05, 07 and 09). At certain times, high congestion levels can be observed also on the edge of the platform, for example when the passengers may want to check whether the train is arriving. Then the LOS maps updated in real time provide invaluable information to warn passengers from falling off the platform.

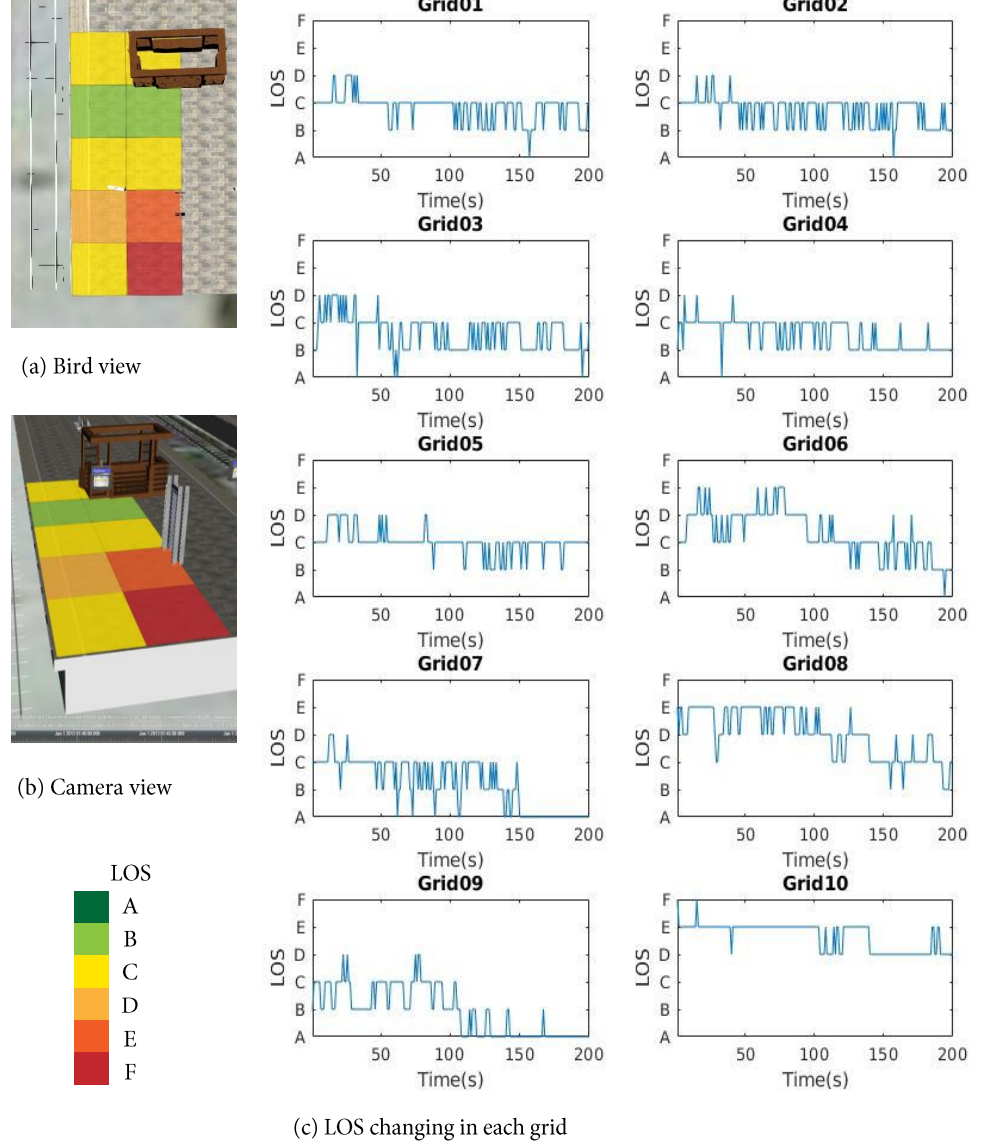


Figure 10: Temporal variation of level of service in each grid

3.4 SUMMARY AND CONCLUSIONS

In this study, a novel approach to direct classification of level of service using CCTV images is proposed, implemented and tested with applications in congestion management of major transport stations. The direct one-step methods avoid complex calculation of detailed flow parameters and greatly improve the detection efficiency therefore achieve real-time monitoring.

Results of an experiment in Flinders Street Station, Melbourne, Australia show that congestion levels can be accurately identified with a total accuracy of 81.9%. It was also shown that exploiting the temporal correlation between successive frames results in

more plausible level of service maps with fewer temporal fluctuations. The proposed approach achieves near real-time performance with an average computation time of 0.40 seconds per frame. Our results demonstrate that LOS maps updated in real time provide invaluable information for monitoring crowd congestion in transport stations and improving the safety of passengers.

MULTI-VIEW LEVEL OF SERVICE MAP GENERATION

This chapter is based on two articles: one peer-reviewed conference paper presented at *Transportation Research Board Annual Meeting* [Li et al., 2019b] and one manuscript entitled "Multi-view crowd congestion monitoring system based on an ensemble of CNN classifiers", submitted to *Journal of Intelligent Transportation Systems* for peer review. As the first author, I was the main researcher proposing the research idea, designing and implementing the experiment, manipulating the datasets, analyzing the results, doing discussions, and writing the manuscript. My supervisors Prof. Majid Sarvi, Dr. Kourosh Khoshelham and Dr. Milad Haghani were regularly discussing with me the research idea, design and progress.

4.1 INTRODUCTION

After generating the congestion map for single camera views, it is necessary to merge the congestion maps of all camera views to obtain a big picture for the entire station. By utilizing complementary information captured by multiple cameras, the issue of limited views and occlusion in single views can be addressed to gain better insight into the whole monitored space. However, multi-view surveillance has been widely applied to microscopic crowd analysis, for example pedestrian detection and tracking, while macroscopic level analysis, which deals with the whole crowd, has received little attention.

In this study, we propose a multi-view framework for the generation of level of service maps, which are the most commonly used measure of congestion at macroscopic level, based on an ensemble of state-of-the-art Convolutional Neural Networks (CNNs). The contributions of this research are two-fold:

1. First, a multi-view crowd congestion map fusion framework is proposed for macroscopic crowd congestion management. Comparison of multiple classifier combination techniques and evaluation on two datasets show that the model can not only improve the performance of crowd density map gener-

ation but also address the detection difficulties in occluded views.

2. Second, a 3D GIS platform is developed to visualize the crowd density maps in real-time and a loading test is conducted to test the maximum number of cameras for which the proposed multi-view congestion classification method can be run in real-time, which determines camera coverage area and is useful for practical application of implementing multi-view crowd monitoring. The results of a loading test show that a maximum of 48 cameras can be processed at a map refresh rate of 2 seconds.

The rest of the article is organized as follows. The framework of multi-view level of service map fusion is introduced in Section 4.2. Section 4.3 describes the experiments and results on two datasets, both in sparse and dense scenarios. Finally, the conclusions are summarized in section 4.4.

4.2 MULTI-CAMERA FUSION FRAMEWORK

As shown in Figure 11, the multi-view level-of-service map fusion system consists of four layers: back-end data layer, middle processing layer, front-end application layer and data transmission layer. The data layer comprises both the CCTV video capture system and video datasets. The processing layer is to process the videos of each camera on cloud servers or local computers and parallel computing is conducted to run multiple processing at one time to improve calculation efficiency. After fusion LOS values from multiple cameras, the final LOS values are generated for application layer. The application layer renders the LOS maps on 3D GIS platform Cesium, which can be accessed by web browser without complex app installation. The LOS maps are generated dynamically and in real-time. Data transmission layer is to transfer data between previous three layers by Wi-Fi, wired network or data input and output modules.

The proposed framework consists of the following key steps: first defining a grid with an appropriate cell size on the ground level. The extent of each grid cell is then projected to each camera view and a level of service label is computed based on the image content. The level of service labels from different views are then fused based on ensemble rules to obtain the final level of service

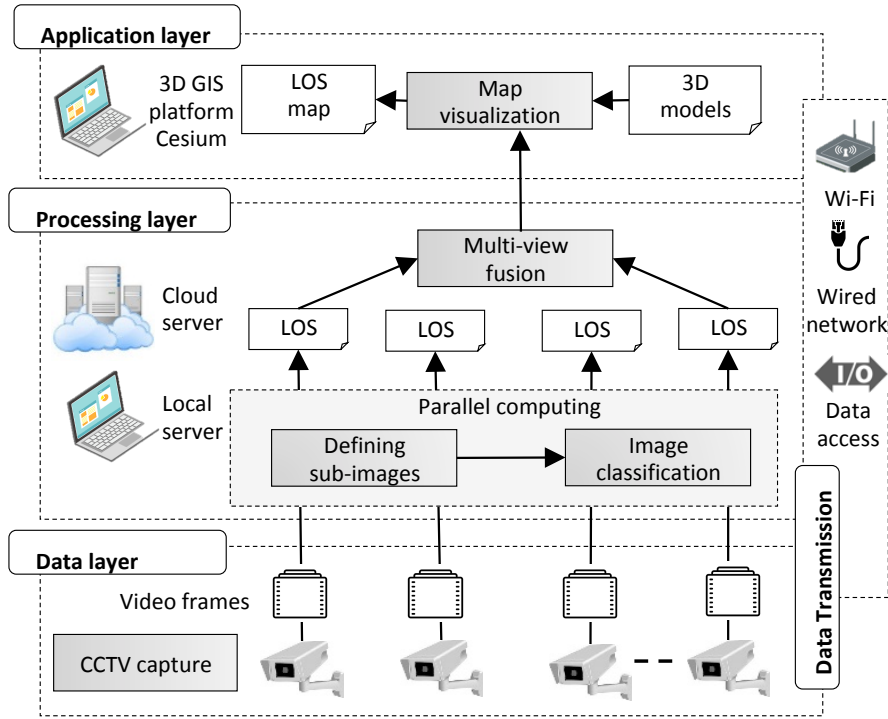


Figure 11: System architecture of multi-view crowd congestion map fusion system

map and visualize it on 3D GIS platform. The algorithms are explained as follows:

4.2.1 Defining sub-images corresponding to the grid cells

We define a congestion map as a grid with appropriate cell size on the ground plane. This representation is very useful to show the distribution of congestion, especially the congested places, the basic unit of fusion from multiple cameras is a grid cell. An illustration of defining the grid cells and back projecting the cells to the camera views is displayed in Figure 12. Firstly, the map of the monitored place is divided into several grid cells. In crowd dynamics, the grid area is usually defined as approximately 10 m^2 . Lo and Velastin, 2001 allocates one level of service to the entire camera view, which covers a broad area, larger than $20\text{m} \times 20\text{m}$ and results in a low-resolution level of service map. Considering the practical needs of congestion monitoring and the size of monitoring area in our study, which is $75\text{m} \times 50\text{m}$ and $15\text{m} \times 15\text{m}$, we set the grid unit to $2\text{m} \times 2\text{m}$ in this research. Secondly, each grid cell is projected to each camera view using the homography matrix. The homography matrix represents the transformation be-

tween the corner points of the grids in camera planes and the ground plane.

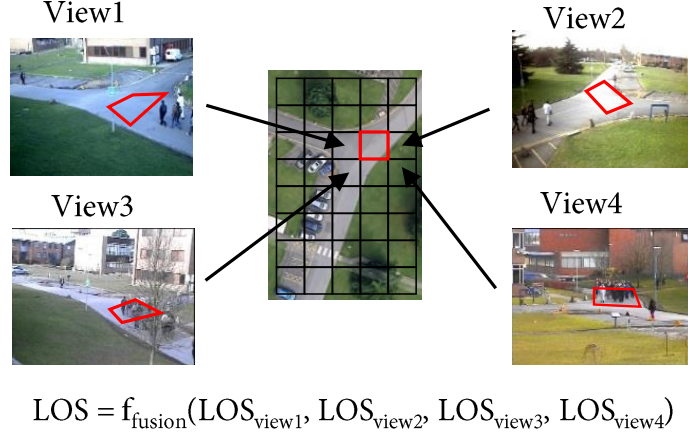


Figure 12: An illustration of define sub-images corresponding to the grid cells.

In order to estimate the coefficients of homography matrix in each view, four pairs of corresponding points are chosen. We denote the camera view by k and n is the n_{th} image point we choose from each camera view. Let $(x_{kn}, y_{kn}, 1)$ denote the homogeneous image coordinates of the points in view k , and $(X_{kn}, Y_{kn}, 1)$ denote the world coordinates of the image points. The coordinates in the camera view are in pixels and the point coordinates in the ground plane are in meters. In order to obtain world coordinates, the image points are labeled on Google Maps, where the longitude and latitude can be read. Then Universal Transverse Mercator (UTM) projection is conducted to convert the longitude, latitude coordinates to x, y coordinates in meters. The mapping from the image coordinates to the world coordinates is a 3×3 homography matrix denoted as H , as shown in equation (1).

$$\begin{bmatrix} X_{kn} \\ Y_{kn} \\ 1 \end{bmatrix} = H \times \begin{bmatrix} x_{kn} \\ y_{kn} \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \times \begin{bmatrix} x_{kn} \\ y_{kn} \\ 1 \end{bmatrix} \quad (1)$$

4.2.2 Single view level of service classification

Our previous work (Li et al., 2019a) has demonstrated the effectiveness of combining a Convolutional Neural Network (CNN) with the Support Vector Machine (SVM) to generate level of ser-

vice maps in single camera views. As shown in Figure 13, we feed every sub-image corresponding to a grid cell to a CNN for feature extraction. The state-of-the-art deep learning model VGG-VD model (Simonyan and Zisserman, 2014) along with pre-trained weights on ImageNet database is applied to extract the features. The original CNN model is composed of 13 convolutional layers and 3 fully connected layers. As the SVM shows excellent performance in high-dimensional classification problems with insufficient training samples, we replace the last layer of the original CNN structure (softmax layer) with the SVM classifier. We cross-validate the SVM model for LOS classification and the Radial Basis Function (RBF) kernel works best for our dataset.

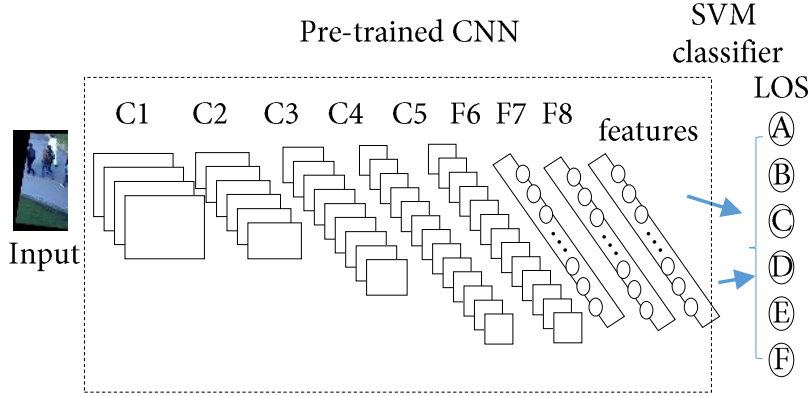


Figure 13: Single view level of service classification based on a Convolutional Neural Network and Support Vector Machine. LOS A to F is equal to 0-1, 2, 3, 4, 5-8, 9 or more pedestrians per grid with a grid cell size of $2\text{m} \times 2\text{m}$, respectively.

4.2.3 Merging level of service at different views

Using the above procedures, for each grid cell we obtain a soft level of service label (i.e. a probability) from each camera view. To obtain a single level of service for the grid cell, we need to combine the different level of service labels from different views. Three commonly used combination rules include algebraic methods, voting methods and borda count (Zhang and Ma, 2012).

Assuming j is one level in the universal set of all levels of service, $\{A, B, C, D, E, F\}$, k is the order of camera views, and the capital K is the total number of cameras, the decision-function method of SVM can give per-class probabilities for each image, denoted as $P(\text{LOS}_j | \text{view}_k)$. The sum of all the probabilities in one view is 1.

4.2.3.1 Algebraic method

Algebraic methods combine the individual probabilities obtained by each class through an algebraic expression, such as maximum, sum, mean, product, etc.. The final ensemble decision is the level that receives the largest probability, as formulated in equation (2).

$$\text{LOS}_{\text{final}} = \text{argmax}_j u_j(P(\text{LOS}_j|\text{view}_k)) \quad (2)$$

The algebraic expressions, denoted as $u_j(P(\text{LOS}_j|\text{view}_k))$, are computed as follows:

Sum rule: The probabilities of each LOS in all camera views are added, and the LOS with the highest sum of probabilities is chosen as the final choice for the grid cell.

$$u_j = \sum_{k=1}^K P(\text{LOS}_j|\text{view}_k) \quad (3)$$

Mean rule: The mean rule is equal to sum rule over every possible value weighted by the probability of that value. Here the probability of each LOS value is equal, recorded as $\frac{1}{K}$.

$$u_j = \frac{1}{K} \sum_{k=1}^K P(\text{LOS}_j|\text{view}_k) \quad (4)$$

Product rule: The direct product of the probabilities of each LOS in all camera view is calculated. The LOS with the largest product of probabilities is selected as the final LOS.

$$u_j = \prod_{k=1}^K P(\text{LOS}_j|\text{view}_k) \quad (5)$$

Maximum rule: The rule returns the largest probability of each LOS within all camera views. The LOS with the largest maximum probability is selected as the final LOS.

$$u_j = \max_{k=1, \dots, K} P(\text{LOS}_j|\text{view}_k) \quad (6)$$

4.2.3.2 Voting method

Voting methods work on the LOS only, and do not involve with probabilities of each LOS. They choose the LOS that has a majority, that is, receiving the largest total vote, which usually includes

majority voting and weighted majority voting. Weighted majority voting is to give a positive weight to the classifiers in the pool. Since in this paper, the SVM classifiers are applied to each camera view, we assume that they have same weights.

Majority voting rule: Let $C(\text{LOS}_j|\text{view}_k)$ denotes the vote of LOS in each camera view, the value is one when receiving one vote, and zero with no vote. The votes in all camera views are added. The LOS with the largest votes is chosen.

$$\text{LOS}_{\text{final}} = \max_{j=A,\dots,F} \sum_{k=1}^K C(\text{LOS}_j|\text{view}_k) \quad (7)$$

4.2.3.3 Borda count

Borda count not only considers the LOS with the highest votes, but also the one with the second highest vote, which ranks order the LOS by giving each candidate a number of points corresponding to their votes. Assuming variable p is the points based on ranking and q is the ranking order, a candidate receives $q - 1$ points for a first preference, $q - 2$ for a second, etc., with zero points for being ranked last (or left unranked).

$$\text{LOS}_{\text{final}} = \max_{j=A,\dots,F} (p \times \sum_{k=1}^K C(\text{LOS}_j|\text{view}_k)) \quad (8)$$

Table 8 is an example listing the results of all combination rules. The classifier gives the probabilities of each label in each view. The max probabilities (bold numbers in the table) are the final choices for each algebraic rule. The voting column lists the votes each level receives. The borda count column lists the scores for each level based on the votes. The level with the highest votes or scores is the final choice.

4.2.4 Level of service map visualization

In our previous work (Li et al., 2019a), we developed a web-based platform for crowd congestion management. Heatmaps overlaid on a 3D model provide more accurate monitoring of the congestion in large public spaces, especially multi-storey buildings. In this research, we improved the platform capabilities and added multi-view functions. As shown in Figure 14, the level of service

LOS	Unimodal				Algebraic				Voting	Borda
	View1	View2	View3	View4	Sum	Mean	Product	Max	Majority	Borda
A	0.892	0.886	0.080	0.287	2.145	0.536	0.018	0.892	2	4
B	0.04	0.061	0.195	0.445	0.741	0.185	2.12E-04	0.445	1	1
C	0.027	0.022	0.631	0.234	0.914	0.229	8.77E-05	0.631	1	1
D	0.023	0.021	0.051	0.017	0.112	0.028	4.18E-07	0.051	0	0
E	0.014	0.007	0.031	0.011	0.063	0.016	3.34E-08	0.031	0	0
F	0.004	0.003	0.012	0.006	0.025	0.006	8.64E-10	0.012	0	0
Decision	A	A	C	B	A	A	A	A	A	A

Table 8: An example of deciding the final level of service based on several combination rules

maps of all the cameras are visualized on the platform. A camera list is provided for the user to choose the camera view and control the visibility of the level of service maps. The field of view (FOV) of each camera are outlined by colored lines. A video player window can be open to play current CCTV video the user selects. Videos from different camera views are synchronized to display the overall LOS map in real-time.

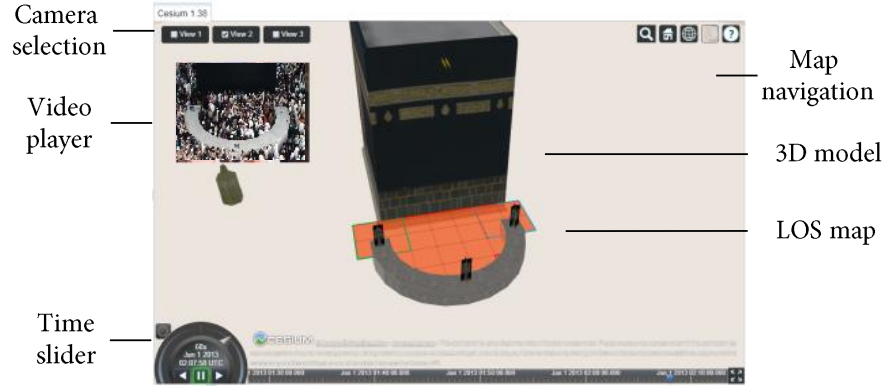


Figure 14: The overall LOS maps is visualized in a web-based interface. The fields of view of each camera are outlined by colored lines.

4.3 EXPERIMENTS AND FINDINGS

We evaluate the performance of the proposed framework using two datasets representing both sparse and dense scenarios. First, we describe the experimental settings, then discuss the accuracy

and efficiency results of the method. Comprehensive analysis on some components of this framework is also presented.

4.3.1 Datasets

In order to sufficiently test on both sparse and dense scenarios, two datasets are used in this paper, namely public dataset PETS2009 (Ferryman and Shahrokni, 2009) and a dataset obtained by our research group from Mecca.

PETS2009 dataset is a public dataset, as shown in Figure 15(a). The cameras are installed at four locations to monitor the road intersection area covering an approximate area of $75\text{m} \times 50\text{m}$. The camera views are projected to the bird view map after computing the homography matrix between adjacent camera views, which is displayed in the right of Figure 15(a). This dataset includes three movements and each movement contains 221 frame images. The levels of pedestrian flows covers light and medium, from LOS A to D.

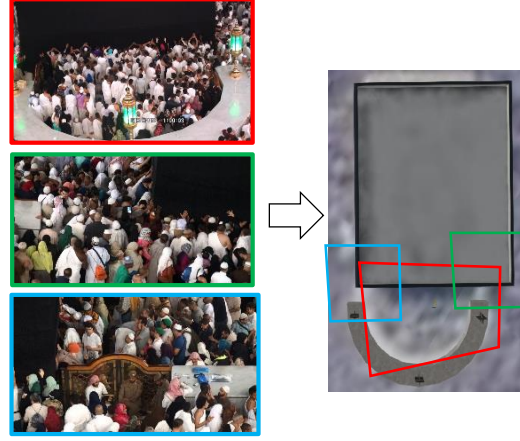
Mecca dataset is collected during Hajj in the city of Mecca, Saudi Arabia. Millions of people pray around the central Kabba as shown in Figure 15(b). The 10 minute video clip records the period when people quickly enter the empty Kabba and fully occupy the place, which mainly covers high level of service from D to F. The cameras are installed at three directions, including west, east and south. The area of central place is approximate $15\text{m} \times 15\text{m}$. The camera views are shown in the left of Figure 15(b). The corresponding fields of view of these cameras are shown in the right of Figure 15(b).

As the datasets are recorded by multiple cameras, video synchronization is required to extract all multi-view frames at certain time. Our approach is to recover the timestamp for all generated frames. For PETS2009 S1-L1 dataset, the video frames are organized and named by time stamps, all the frames at the same time stamp can be easily searched and associated by the file name. For Mecca datasets, the cameras record the start recording date and time, and then the timestamp of all frames can be inferred by the starting time and frame rate.

In order to generate the ground truths, we label the number of pedestrians in each grid. Level of service is defined as the average occupancy area of each person, which means dividing the area of grid by the number of pedestrians. According to this standard, the



(a) PETS2009 dataset



(b) Mecca dataset

Figure 15: The camera views are projected to the ground view plane after computing the homograph matrix for all camera views.

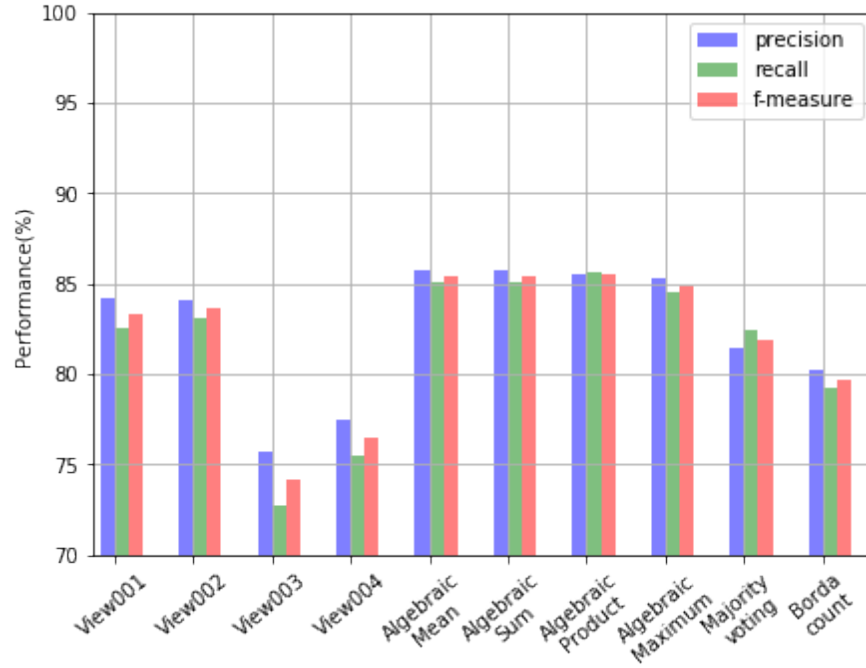
labeling level of service for each grid can be calculated. In order to save the human labor of labeling, for less crowded scenes, we initially apply Mask R-CNN (He et al., 2017) to detect and segment pedestrians in the images, then calculate the number of human shapes in each grid. The labeling results are verified visually. For more crowded scenes, the locations of pedestrians are manually labeled using the labeling tool developed in this research and another labeling tool developed by Idrees et al., 2013. To reduce the errors caused by camera setting angles and individual perspectives, every camera views is labeled by three persons. The label with the highest votes is selected as the final decision. For the PETS2009 dataset, only the part S1-L1 was labeled, which contains 221 frames and 5,525 grid images. For the Mecca dataset, an 8-minute segment of the video was labeled, which included 480 frames and 8,640 grid images.

4.3.2 Comparison of ensemble rules

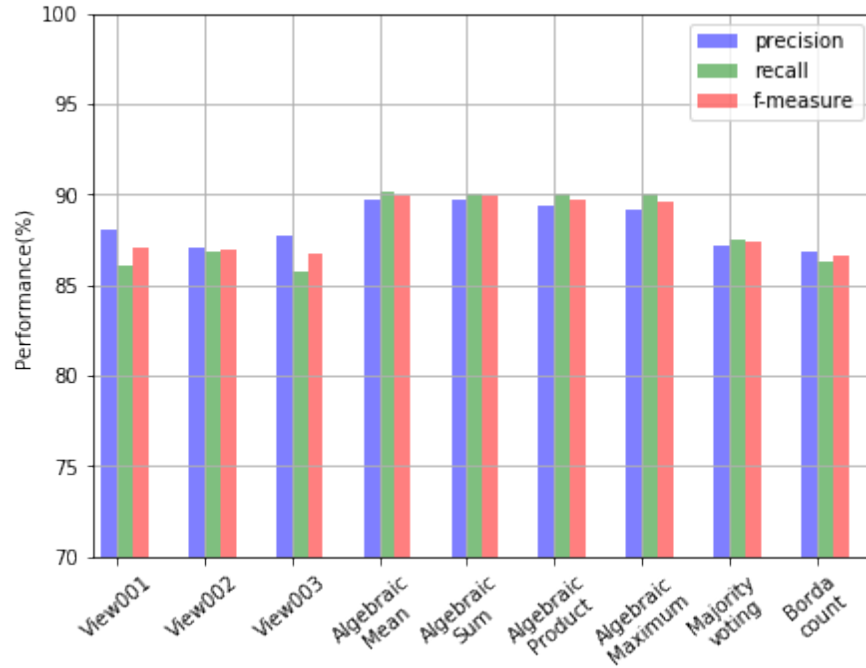
Several ensemble rules are first examined over the two datasets based on the rules in Section 2.3. Three evaluation metrics are used to evaluate the performance of the method: precision, recall and F-measure. Precision is the number of true positives divided by the sum of true positives and false positives. Recall, also known as sensitivity, is the number of true positives divided by the sum of true positives and false negatives. F-measure is defined as the harmonic mean of precision and recall.

As seen in Figure 16, the performances of algebraic methods are relatively higher than voting method and borda count. Compared to algebraic methods, voting method and borda count consider only the votes and ignore the probabilities of each LOS, hence miss some information. Furthermore, when several levels receive the same number of votes, it is hard to decide which one is the best choice. Among algebraic methods, sum rule and mean rule show the same performance as we suggest in Section 2. However max rule is easily affected by occluded and low resolution views, for example the view 003 and 004 in PETS2009 dataset, where a tree occupies most part of the camera view and the branches interfere the human detection; See Figure 15. When the highest probability appears in these views, the max rule would recognize this label as the final decision and cause errors. Other algebraic rules except max rule show better performance because they require all camera views have a high accuracy for this level so that the final result for this level could be high after algebraic operations. If only one view shows a high accuracy for this level, it can not represent the final result.

For the two datasets with different levels of crowdedness, the performances per LOS label obtained by the algebraic sum rule are shown in Table 9 and Figure 17. The sum rule based model yields 86.2% to 89.8% accuracy for level of service classification. Overall, the framework shows better performance on more dense scenarios, as we see from the confusion matrix, where the overall accuracy of Mecca dataset is higher than that of PETS2009 dataset, which has less crowded scenes than Mecca dataset. In the PETS2009 dataset the classification accuracy is higher for LOS A to D, while in the Mecca the classification results are better for LOS C to F. The difference in the classification results can be explained by the imbalance of samples between different LOS labels.



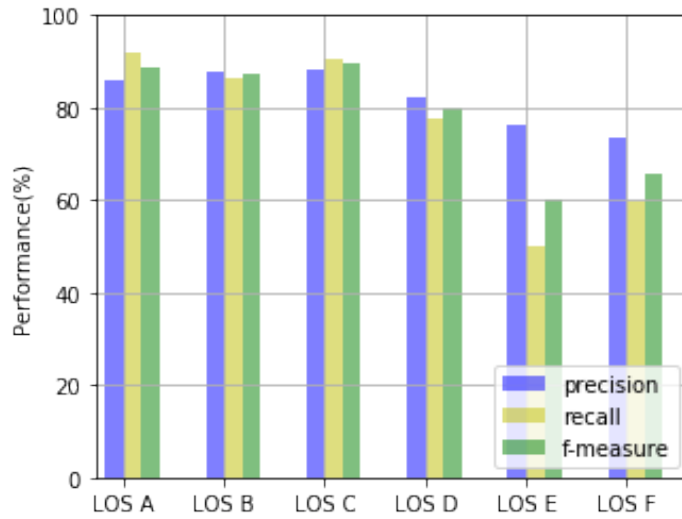
(a) PETS2009 dataset



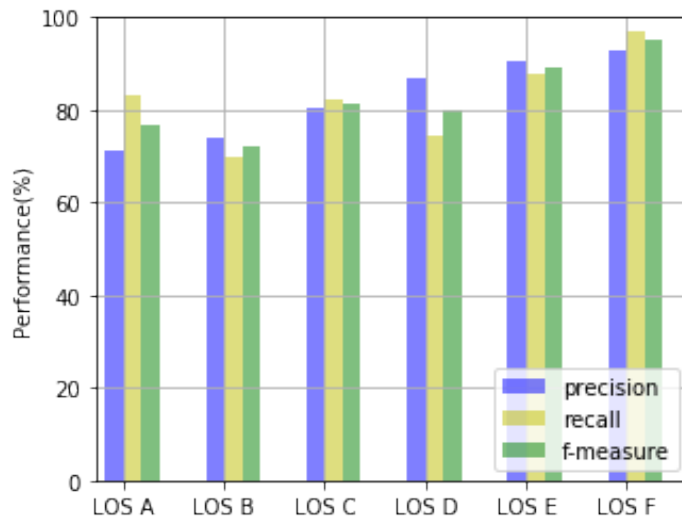
(b) Mecca dataset

Figure 16: Comparisons of ensemble learning techniques. As seen, ensemble learning techniques can help improve classification accuracy. Algebraic based methods show better performance than other techniques.

The fusion of level of service derived from multiple views can improve the performance when one view is occluded. In our experiments, view003 in PETS2009 dataset is occluded by a tree. The



(a) PETS2009 dataset



(b) Mecca dataset

Figure 17: Performance of algebraic sum rule. PETS2009 dataset shows better performance in low LOS (from A to D), and Mecca dataset has higher accuracy in high LOS (from C to F).

detection result of this single view is quite low, only 76.8%, as displayed in Figure 17. However, after merging all camera views the detection accuracy significantly improves and reaches 87.9%. This comparison shows that the settings of multi-view cameras can greatly improve the accuracy of level of service map generation, thus improve vision around blind spots in one view.

	PETS2009						Mecca					
	A	B	C	D	E	F	A	B	C	D	E	F
	1657	1663	1477	512	122	94	377	438	733	814	2113	4165
Prediction												
LOS A	1422	49	22	2	1	1	269	29	12	7	1	6
LOS B	138	1458	76	9	2	3	66	324	49	12	2	15
LOS C	44	97	1302	18	1	3	24	47	588	32	10	10
LOS D	34	23	55	421	3	7	12	23	67	707	121	91
LOS E	15	24	15	50	93	11	3	8	15	45	1916	169
LOS F	4	12	7	12	22	69	3	7	2	11	63	3874
Accuracy	86.2%						89.8%					

Table 9: Confusion matrix using algebraic-sum based method.

4.3.3 Effect of number of views

In the real scenarios, hundreds of cameras may be installed and many cameras would have overlapping views. To investigate whether all camera views are needed in data fusion and how many cameras minimum can achieve an acceptable accuracy, the results from different numbers of cameras were compared. In this experiment, all the combinations of cameras are tested on PETS2009 dataset using the algebraic sum rule, which shows the best performance in the last experiment. As shown in Figure 18, the classification accuracy of individual views are relatively low. In particular, View003 shows the lowest accuracy because of the occlusion by the trees, which makes it hard to recognize individual pedestrians. The overall accuracy of combining all camera views is a bit higher than individual views. The combination of view001 and view002 achieves the best performance. This result shows that not all camera views contribute to improving the detection results. In general, views with occlusion, lower resolution, and poor geometry may deteriorate fusion results and should be removed from the fusion process. Based on the camera settings of this dataset, two camera views with high resolution and no occlusion can achieve the best performance.

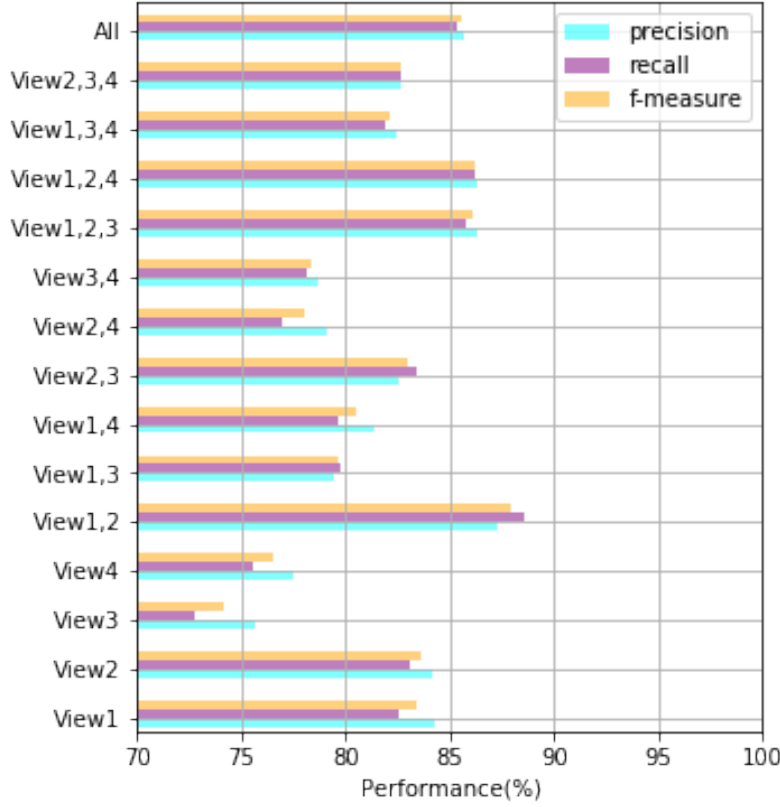


Figure 18: Combinations of cameras are evaluated on PETS2009 dataset using algebraic sum rule. Two camera views with high resolution and no occlusion can achieve optimized performance.

4.3.4 Computational efficiency

A loading test is conducted to test the maximum number of cameras for which the proposed multi-view congestion classification method can be run in real-time. Because maximum four cameras views can be obtained in our dataset (PETS2009 dataset), in order to solve the problem of limited data sources, the processing time of different cameras is measured and compared. The result shows that the difference of processing time between multiple cameras is less than 0.03 seconds, which is negligible. Also in order to quickly reach the limit of the operating system, our loading test method is to add 10 cameras each time. To gain faster overall execution, multi-threading method is used in our framework, which is to put central processing unit in a multi-core processor to execute multiple processors concurrently within the support of the system.

The loading test result is shown in Figure 19. The framework is run on a 2.6-GHz Intel Core i7 processor with 16GB of memory. The processing time for one camera is 0.48 seconds, of which

the image classification step accounts for 0.38 seconds. If the congestion map is required to be refreshed every 1 second, then a maximum of 17 cameras can be processed at one time. If a refresh interval of 2 seconds is acceptable, 48 cameras can be operated.

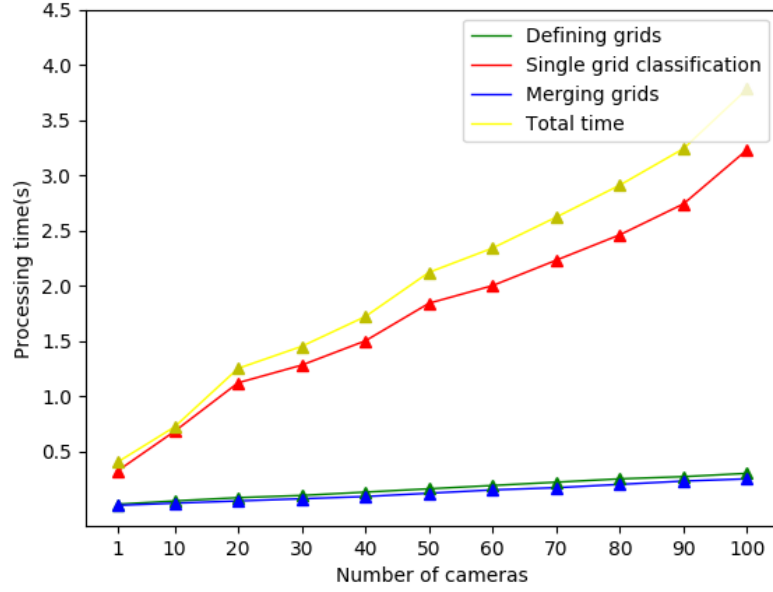


Figure 19: Loading test, which measures the maximum number of cameras executed on this system. In order to achieve a map refresh rate of 2 seconds, maximum 48 cameras views can be run at the same time under the test environment.

4.4 CONCLUSIONS

In this paper, we presented a multi-view fusion framework of generating crowd congestion maps, which provide insights into macroscopic crowd management in large public space. The multi-view fusion methods based on ensemble rules were compared and evaluated. The results show that algebraic methods outperform voting methods and borda count, because algebraic methods consider both the probabilities and the results of all camera views. Two datasets both in sparse and dense scenarios were tested. The final results show that the fusion of camera views can not only provide a wide coverage of monitored space but also help improve the classification accuracy of individual views, especially the occluded views. A loading test was conducted to evaluate the maximum number of cameras on our test environment. With a map refresh rate of 2 seconds, 48 cameras can be processed at the

same time, which means that a density map covering 48 cameras views would be visualized.

For the future work, as the monitored space becomes complex and the size of camera network fast increases, there are higher requirements for the robustness, efficiency, flexibility of the multi-view fusion schema (Wang, 2013). Our experiments prove that the overall accuracy of crowd congestion map generation can be improved by the multi-view fusion framework. More factors can be considered to improve the accuracy, e.g. camera angle, height information. With higher camera angle, the occlusion can be reduced and more people can be seen in the camera view. If height information is added in 3D transformation, the calculation of density map would be more accurate. An efficiency test is also conducted in this paper to test the feasibility of the application. The percentage of overlapping camera views is an important factor in the efficiency test as larger overlapping areas require longer fusion time. It can be seen that the design of camera network is very important as it can not only improve fusion accuracy but also efficiency. Multi-view video surveillance is indispensable for large area crowd congestion management. We believe that our work would encourage new research in this area.

PREDICTING TRAFFIC CONGESTION USING CONVOLUTIONAL LONG SHORT-TERM MEMORY NETWORKS

This chapter is composed based on two papers: one peer-reviewed conference paper of the same title presented in *International Symposium of Transport Simulation (ISTS'18)* and the *International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)* and another journal paper entitled "Predicting traffic congestion using convolutional long short-term memory networks" submitted for peer review. As the first author, I was the main researcher proposing the research idea, designing and implementing the experiment, manipulating the datasets, analyzing the results, doing discussions, and writing the manuscript. My supervisors Prof. Majid Sarvi and Dr. Kourosh Khoshelham were regularly discussing with me the research idea, design and progress.

5.1 INTRODUCTION

Predicting the congestion over a relatively short period of time is of great importance for traffic management and public safety. Previous methods require data from multiple sources and involve multiple parameters, such as traffic flow, volume, occupancy, speed and ramp flow for the analysis of congestion. Congestion prediction is still a challenge due to the variety of data sources required and the complex calculation of traffic parameters. Little research has examined the congestion prediction problem from the image processing perspective, which considers only images as inputs and outputs. The advantage of image-based congestion prediction is that it is direct and hence no intermediate parameters are needed.

In this paper, we propose an innovative spatio-temporal deep learning method for congestion analysis requiring only historical congestion maps to predict future congestion maps. The proposed approach provides insights to understanding the spatial and temporal patterns from traffic images which could greatly simplify

the prediction task. In summary, the main contributions of this research are:

- to predict short-term congestion using only historical and current traffic images, without requiring other traffic parameters.
- to introduce a novel spatio-temporal deep learning method, called convolutional long short-term meomroy (ConvLSTM), to accurately predict future traffic congestion in both free-way and urban roads scenarios.
- to convert the images to GIS data for a wider application of congestion maps in congestion analysis.

The rest of this paper is organized as follows. Section 5.2 presents our approach of congestion prediction. Section 5.3 discusses our experiments in freeway and urban roads scenarios. Concluding remarks and future work are described in Section 5.4.

5.2 METHODOLOGY

5.2.1 Problem statement

The goal of congestion forecasting is to use previously observed traffic image sequences to forecast the future congestion maps in a certain period of a specific region. Suppose we observe the traffic over a spatial region represented by an $M \times N$ grid which consists of M columns and N rows. Inside each cell in the grid, the congestion level varies over time. Thus the observation at any time can be represented by a 2D tensor $T \in \mathbf{R}^{M \times N}$, where $\mathbf{R}$ denotes the domain of observed traffic maps. Since we record the traffic maps periodically, we get a sequence of 2D tensors, which we can concatenate into a 3D tensor $\hat{T}_1, \hat{T}_2, \dots, \hat{T}_t$, as shown in Figure 20. The spatio-temporal sequence forecasting problem is to predict the most likely K length sequence in the future given the previous J observations which include the current status:

$$\hat{T}_{t+1}, \dots, \hat{T}_{t+K} = \arg \max_{T_{t+1}, \dots, T_{t+K}} p(T_{t+1}, \dots, T_{t+K} | \hat{T}_{t-J+1}, \hat{T}_{t-J+2}, \dots, \hat{T}_t) \quad (9)$$

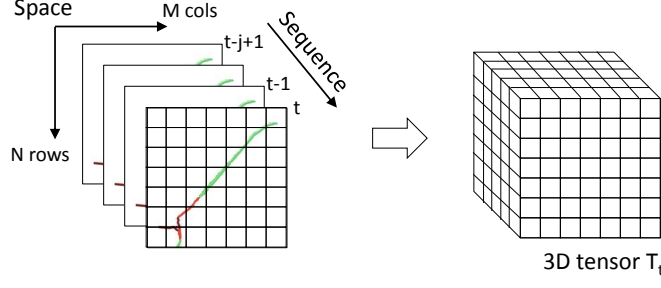


Figure 20: Input data formulation

5.2.2 Network architecture

5.2.2.1 ConvLSTM model

We use the convolutional long short-term memory (ConvLSTM) model Xingjian et al. (2015), which extends the fully connected LSTM layer to have convolutional structures in both the input-to-state and state-to-state transitions, so that it can learn spatial and temporal dependencies at the same time .

The ConvLSTM model still adopts the input gate, forget gate, output gate and memory cell in the traditional fully connected LSTM structure in order to control the passage of information at each time step so that the long-term temporal dependencies in the traffic maps can be captured. The memory cell C_t acts as an accumulator of the cell state information. When a new input T_t comes, its information will be accumulated to the cell if the input gate i_t is activated, and the past cell status C_{t-1} would be forgotten if the forget gate f_t is activated. Finally the output gate o_t determines whether the cell status C_t would be propagated to the hidden state H_t also known as the output vector of each ConvLSTM unit.

All the inputs $T_1, \dots, T_t$, cell outputs $C_1, \dots, C_t$, hidden states $H_1, \dots, H_t$ and gates i_t, f_t, o_t of the ConvLSTM model are 3D tensors whose last two dimensions are spatial dimensions (rows and columns), so that spatial information can be encoded. The ConvL-

STM model generates the next cell state C_t and hidden state H_t by the following steps Xingjian et al. (2015):

$$\begin{aligned}
 i_t &= \sigma(W_{xi} * T_t + W_{hi} * H_{t-1} + b_i) \\
 f_t &= \sigma(W_{xf} * T_t + W_{hf} * H_{t-1} + b_f) \\
 o_t &= \sigma(W_{xo} * T_t + W_{ho} * H_{t-1} + b_o) \\
 g_t &= \tanh(W_{xc} * T_t + W_{hc} * H_{t-1} + b_c) \\
 C_t &= f_t \circ C_{t-1} + i_t \circ g_t \\
 H_t &= o_t \circ \tanh(C_t)
 \end{aligned} \tag{10}$$

where $\{W_{xi}, W_{hi}, b_i\}$, $\{W_{xf}, W_{hf}, b_f\}$, $\{W_{xo}, W_{ho}, b_o\}$, and $\{W_{xc}, W_{hc}, b_c\}$ are the parameters of the input gate, forget gate, output gate and memory cell, respectively. σ is the sigmoid function, $\tanh$ is the hyperbolic tangent function, $*$ denotes the convolution operator and $\circ$ denotes the Hadamard product.

5.2.2.2 Prediction structure

The prediction structure proposed in this research is composed of four ConvLSTM layers and one 3D convolution (Conv3D) layer, as illustrated in Figure 21. Each ConvLSTM layer contains 40 hidden states and 3×3 kernel. The Batchnormalization layers normalize the values of the previous layer at each batch and apply a transformation that maintains the mean activation close to 0 and the activation standard deviation close to 1. The last Conv3D layer creates a 3D convolution kernel convolved with the layer input over the temporal dimension to produce time series outputs. The kernel size of the Conv3D layer is $3 \times 3 \times 3$.

5.2.2.3 Training algorithm

The training process is to fine tune the parameters of the ConvLSTM based prediction structure. The model is trained end to end. The Adam optimizer with a learning rate of 0.001 is applied to minimize the cross entropy loss of the accuracy for the predicted sequences and the true sequences.

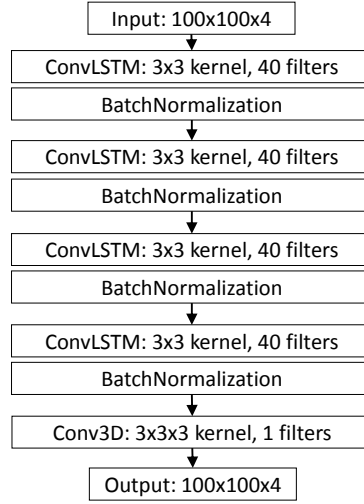


Figure 21: ConvLSTM model based prediction structure

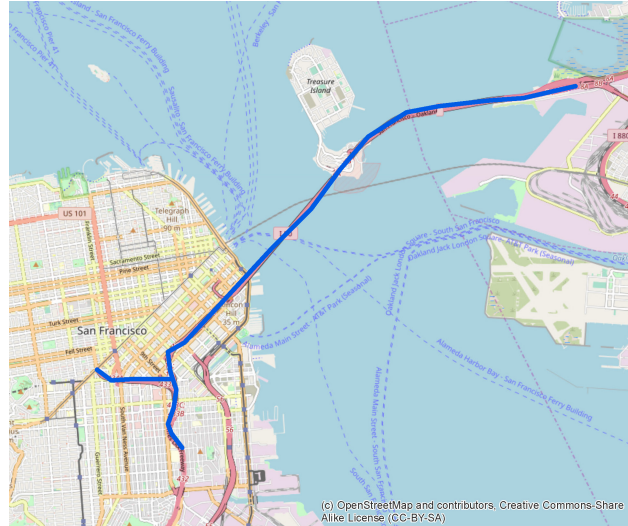
5.3 EXPERIMENTS

5.3.1 Datasets

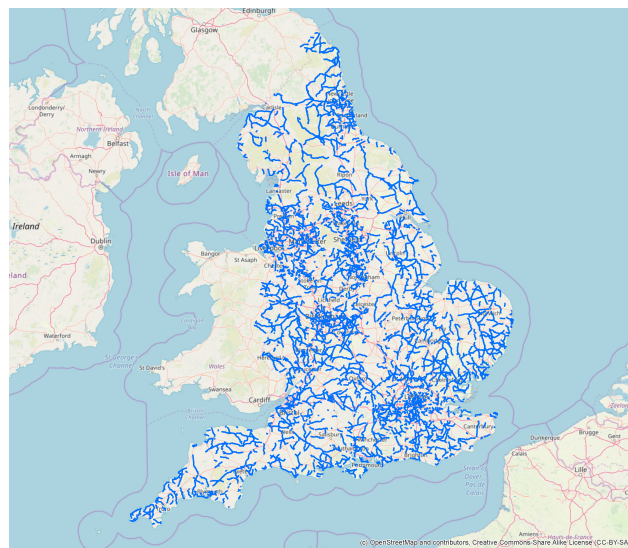
We collected and processed two datasets from the California Department of Transportation and the United Kingdom Department for Transport in order to evaluate the performance of the method on freeways and urban minor roads, respectively.

PEMS I80 freeway dataset (PEMSI80), as shown in Figure 22a, was collected on the freeway I80 in San Francisco, from the Highway 101 to the Treasure Island portion of the Bay Bridge, which is the most congested segment in San Francisco during afternoon peak hours on weekdays. We recorded congestion maps along the freeway every 5 minutes from 19:00 to 20:00 pm on weekdays. This resulted in 12 frames of size 400×400 pixels per day. The data instances were sliced from these frames using an 8-frame-wide sliding window (4 frames for the input and 4 frames for the prediction), thus we get 5 sequences per day and all the sequences are 8 frames long. The time range of the dataset is from 14 October 2014 to 4 February 2018, covering 3 years, in total 866 days, excluding the days that were missing data. We select the two years data, i.e. 2880 sequences, as the training dataset and one year data, i.e. 1450 sequences, as the testing dataset.

UK minor roads datasets (UKMINR) was collected from the United Kingdom Department for Transport and includes all the 1957 minor roads in the major metropolitan areas of United Kingdom, as shown in Figure 22b. The dataset was recorded every 1



(a)



(b)

Figure 22: Datasets:(a) Freeway I80 in San Francisco. (b) The minor roads in United Kingdom.

hour from 7:00 am to 18:00 pm, so there are 12 frames per day. Using the same 8-frame-wide sliding window as the PEMS180 dataset, we get 5 sequences per day. The time range of the UKMINR dataset is the weekdays from 2016 to 2018. We select 1200 minor roads as the training dataset and the remaining 757 minor roads served as the test dataset separately. Excluding the days missing data, the UKMINR dataset contains 56370 training sequences and 28185 testing sequences.

5.3.2 *Data pre-processing*

In the PEMS180 dataset, the complete traffic maps were recorded from the website. A program based on Javascript was developed to automatically run the capture of traffic maps every 5 minutes. Since the aim of this research is to predict the congestion of the roads, other land areas are not of interests. Therefore we extracted the regions of interest by comparing the base maps with the congestion maps and finding the difference of these two images, which is our target road area. Another reason for this step is to save the computing resources by excluding the useless pixels from the calculation. As millions of parameters are calculated in deep learning, in order to save computing resources, the maps are divided into small grids. Based on our experiment, 100×100 pixels grids are the most appropriate image size to be fed into the network in our test environment. Since the outputs of the prediction are traffic layers without base maps. Reconstructing congestion maps can be done by merging the grids and overlaying base maps, so that we can generate the final complete congestion maps.

In the UKMINR dataset, we generate the traffic images based on the raw traffic count data. Firstly we divide the traffic count number into four levels and define the colors for each level. Secondly each minor road are visualized in the ArcGIS platform and scaled to best fit the map size 400×400 pixels. The colors representing each level are assigned to each road links. Then the traffic maps at each time stamp can be produced.

5.3.3 *Experiment settings*

All the experiments are compiled and tested on an Ubuntu cloud server with 110GB memory, 12 virtual CPUs and a NVIDIA Tesla

K80 GPU. The model is implemented in Keras with Tensorflow backend Chollet (2015).

5.3.4 Evaluation metrics

We measure the accuracy of the prediction by calculating the percentage of correctly predicted pixels in the road region:

$$\text{Accuracy} = \frac{\sum_{i=1}^M \sum_{j=1}^N p_{ij}}{M \times N}, \begin{cases} p_{ij} = 1, & \text{if } \tilde{x}_{ij} = x_{ij} \\ p_{ij} = 0, & \text{if } \tilde{x}_{ij} \neq x_{ij} \end{cases} \quad (11)$$

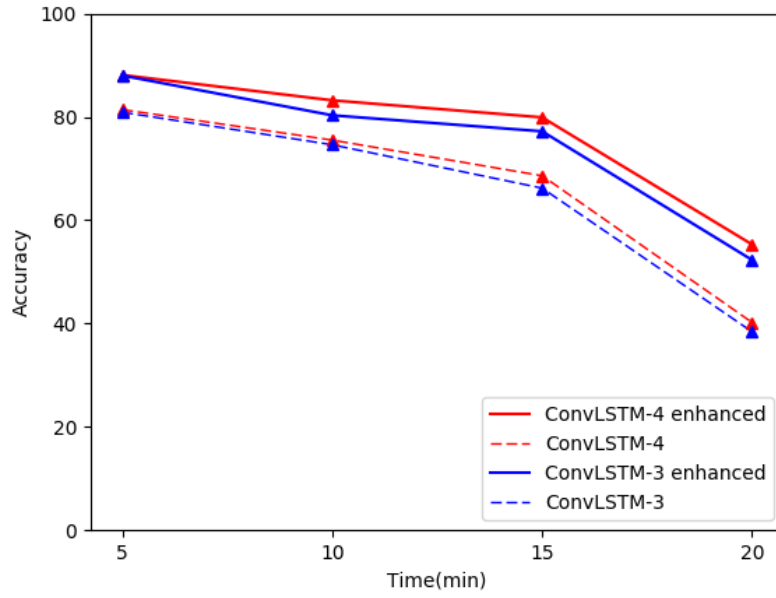
where M and N denotes the total number of columns and rows of the traffic image, i and j denote the i_{th} column and j_{th} row pixel of the traffic image. $\tilde{x}_{ij}$ is the prediction value of the pixel while x_{ij} is the ground-truth of this pixel.

5.3.5 Experiment results

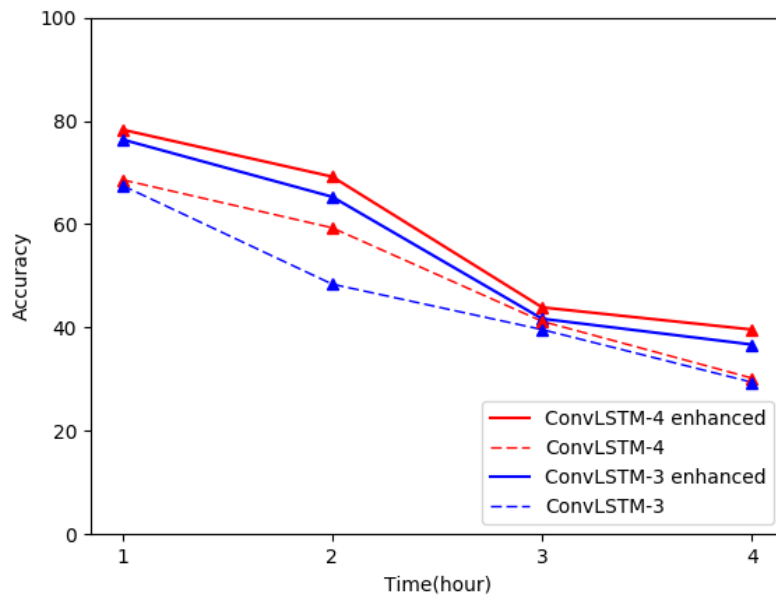
The results of the two datasets are shown in Figure 23, Table 10 and Table 11. In the experiments we compared the prediction structure with 4 ConvLSTM and 3 ConvLSTM layers, respectively. We can find that the deeper structure with 4 layers shows slightly higher accuracy than the less deep structure with 3 layers. In order to solve the blurry problem of the prediction, which can be seen from some examples of prediction in Figure 24, we vectorized the maps and reconstructed the raster maps by retrieving the high proportion values of each road segment and assigning the representing color the road segment again. The results show that the accuracy can also be highly enhanced, which we can observe from Table 23 that the accuracy can be improved up to 15% .

Table 10: The prediction accuracy at different time intervals on PEMS18o dataset

Accuracy(%)	5 mins	10 mins	15 mins	20 mins
ConvLSTM - 4 enhanced	88.1	83.2	79.9	55.3
ConvLSTM - 3 enhanced	88.0	80.3	77.2	52.3
ConvLSTM - 4	81.4	75.5	68.6	40.2
ConvLSTM - 3	80.9	74.6	66.2	38.4



(a)



(b)

Figure 23: Accuracy of each predicted frame before and after enhancement.(a) PEMS18o dataset. (b)UKMINR dataset.

Table 11: The prediction accuracy at different time intervals on UKMINR dataset

Accuracy(%)	1 hours	2 hours	3 hours	4 hours
ConvLSTM - 4 enhanced	78.3	69.2	43.9	39.6
ConvLSTM - 3 enhanced	76.4	65.3	41.7	36.7
ConvLSTM - 4	68.6	59.3	41.2	30.2
ConvLSTM - 3	67.5	48.4	39.6	29.4

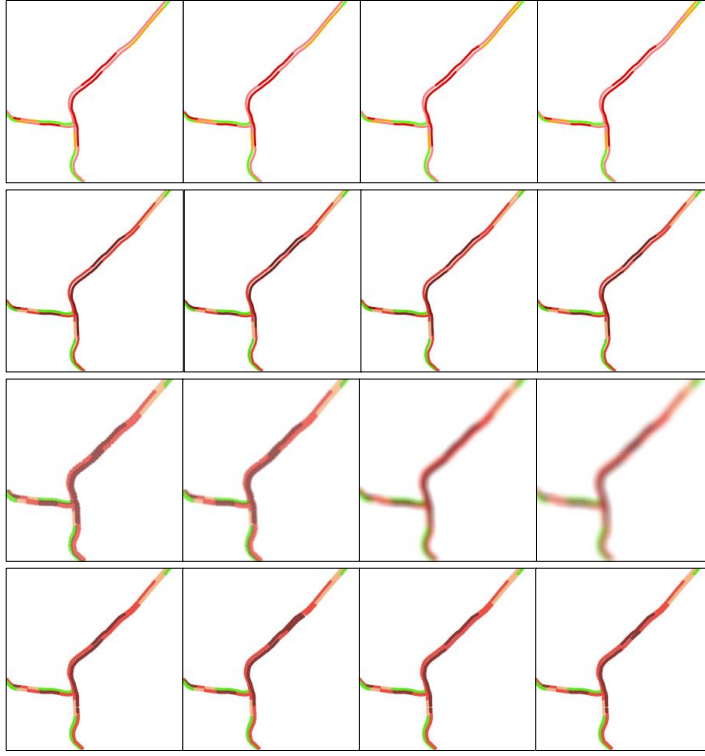
In terms of the prediction time, the accuracy of prediction declines over time. In the PEMS180 dataset, the accuracy of prediction is close to 80% in the first 15 mins. In the UKMINR dataset, the accuracy of prediction is above 75% in the first 1 hour (1 step). Overall, the methodology can achieve higher accuracy in the short-term traffic prediction than the long-term prediction.

In terms of the area of implementation, around three quarters of the research on traffic prediction focus on the expressway, freeway, highway, motorway, carriageway or ring roads Ermagun and Levinson (2018), and only one quarter of the studies use urban roads as the test area. It is noted that the predictions become more complicated in urban area. This is because the freeways usually appear as simple networks or ring networks, while the urban network shows more complicated grid networks. Our results show that the accuracy of prediction for the first future time interval is 88.1% for the highway and 78.3% for the minor urban roads.

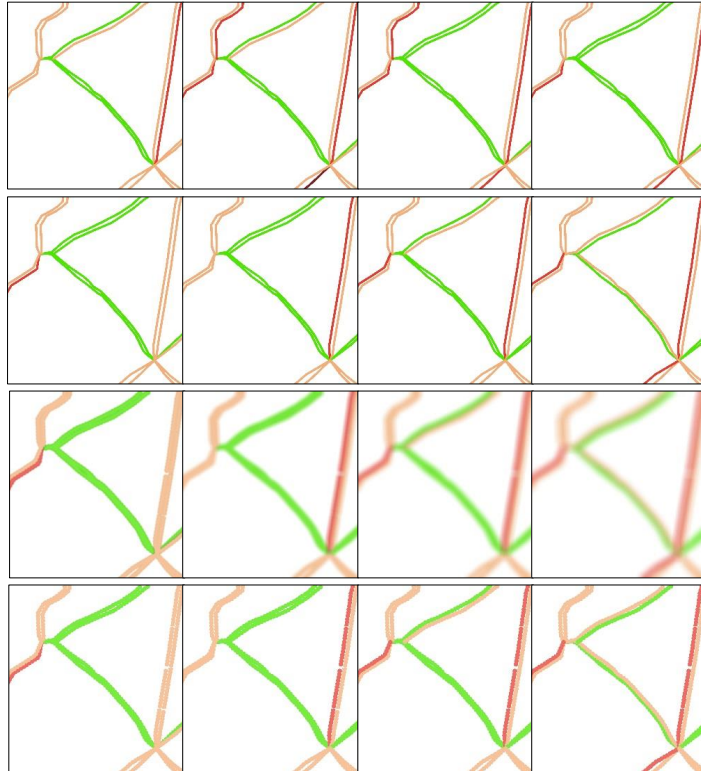
Some prediction examples of the two datasets are shown in Figure 24. We can see that the model can accurately predict the congestion levels of most road segments, but after the convolution filter calculation, the prediction images become blurred. To reduce the blur, we performed map vectorization as we mentioned in the first paragraph. The enhanced images are shown in the fourth row.

5.3.6 Application of congestion length calculation

GIS data with accurate geographic locations and attributes can provide a rich source of information for traffic congestion studies such as travel times and delays, congestion levels, and energy emissions Taylor, Woolley, and Zito (2000). In order to conduct further analysis, we convert the raster congestion maps to GIS data by retrieving the congestion level of each road segment and assign-



(a)



(b)

Figure 24: Some prediction examples, from top to bottom: input frames; ground truth images and prediction images before enhancement, prediction images after enhancement. (a) PEMS180 dataset. (b) UKMINR dataset.

ing the values to the attribute of GIS road objects. Another benefit of this procedure is to deal with the inherently blurry problem of video prediction Mathieu, Couprie, and LeCun (2015). After converting to GIS data, we assign one congestion value to each road segment, which can guarantee that each road segments can have only one congestion level thus improving the accuracy of pixel prediction.

After converting the raster maps to georeferenced maps, we conduct congestion analysis based on the GIS data. The length of traffic congestion is a typical variable showing the present state of congestion, which is defined as the total length of road segments with congestion level above a certain threshold, as shown in Figure 25 in red and brown color. We filter the roads with high congestion level (red and brown colors) and obtain the length attribute of the filtered road, and finally summarize the length value to get the congestion length of the whole road segment. As it can be observed in Figure 26 that the congestion length decreases after 19:00 pm but decreases after 19:20 with 19:30 pm being the most congested time after 19:00 pm.

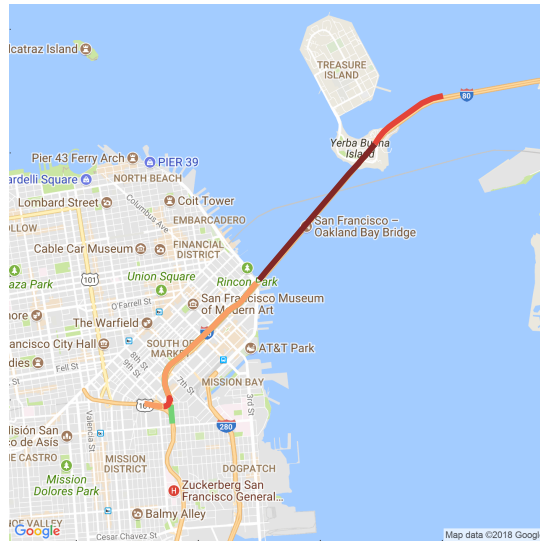


Figure 25: Congestion length of one road, the total of red and brown parts

5.4 CONCLUSION AND FUTURE WORK

In this paper, we applied deep learning method to the challenging problem of short-term congestion prediction, which so far has not been studied from an image analysis perspective based on

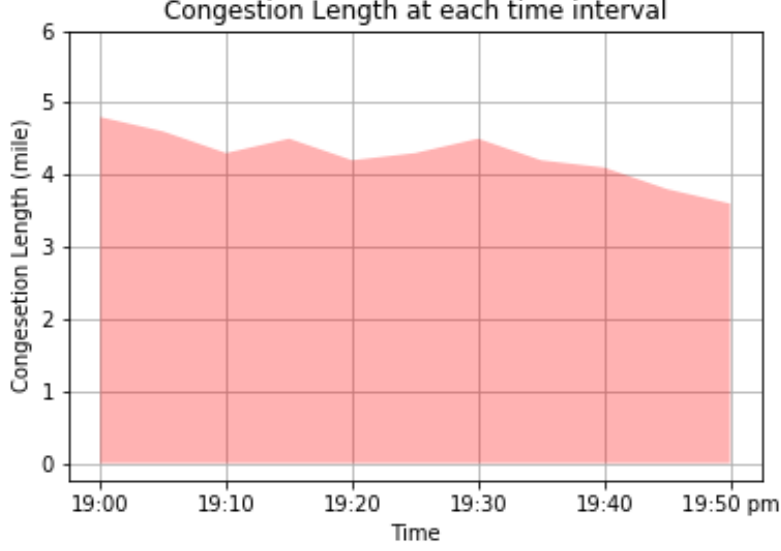


Figure 26: Congestion length by time.

open source congestion maps. We formulate congestion prediction as a spatio-temporal sequence forecasting problem and introduce ConvLSTM structure to tackle the problem. The advantage of our approach is that unlike the previous methods that involve the laborious work of fusing many traffic parameters, our method requires only historical congestion maps for congestion analysis. Based on our experiment on the freeway and urban minor roads, this framework can predict 15 minutes ahead on the highway with an accuracy of 88.1% and 1 hour ahead on the urban minor roads with an accuracy of 78.3%.

We believe that the presented results are promising and this research can stimulate the research on traffic prediction from an image processing perspective, which is very useful for understanding the spatio-temporal dependencies based on realistic road networks. For the future work, our research would continue in several directions: (1) the model can be extended to more traffic parameters such as traffic speed and flow. (2) In order to learn the spatial-temporal dependencies in a wider area, with more computing resources, the depth of the network and the input image size can be increased which can greatly improve the performance of the model.

Part III

MICRO-SCALE PEDESTRIAN ORIGIN-DESTINATION ESTIMATION

Part [iii](#) discusses visual crowd analysis on a micro scale, i.e. analysing the trajectory of individual pedestrians. Chapter [6](#) proposes a pedestrian origin-destination estimation method based on multi-camera person re-identification for evaluating the infrastructure design. Chapter [7](#) applies RGB-D sensors for origin-destination estimation and compares the performance of RGB-D sensors in normal and emergency situations.

PEDESTRIAN ORIGIN-DESTINATION ESTIMATION BASED ON MULTI-CAMERA PERSON RE-IDENTIFICATION

This chapter is adapted from the manuscript of the same title, submitted to *IEEE Internet of Things* for peer review. As the first author, I conducted all parts of the work, including research design and implementation, result analysis, discussions, and paper-writing. My supervisors Prof. Majid Sarvi and Dr. Kourosh Khoshelham were supervising the progress, contributing to research ideas, bringing up critical thoughts and comments at the stage of experiment design, and actively discussing the results.

6.1 INTRODUCTION

Pedestrian origin-destination (O-D) estimation is critical for the management of pedestrian facilities, including pedestrian flow simulation at the planning phase and crowd control at the operation phase. However, current data collection techniques, which include O-D surveys, mobile sensing using GPS, Wi-Fi and Bluetooth, and smart card data, have the disadvantage that they are either time consuming and costly, or can not provide complete O-D information inside the pedestrian facilities without entrances and exits or in the interior of the facilities. To overcome these issues, some studies have explored the image processing techniques, for example people tracking to collect pedestrian trajectories with overhead cameras. But existing image-based pedestrian tracking methods are limited to single camera views and multi-view pedestrian re-identification (re-ID) remains an unsolved problem. To fill in this gap, in this chapter we propose a novel pedestrian O-D estimation framework and address the challenge from three aspects:

1. we develop and test a novel pedestrian O-D estimation framework based on multi-camera pedestrian re-identification (re-ID). By identifying the same person in different camera views, the O-D of each identity can be generated.

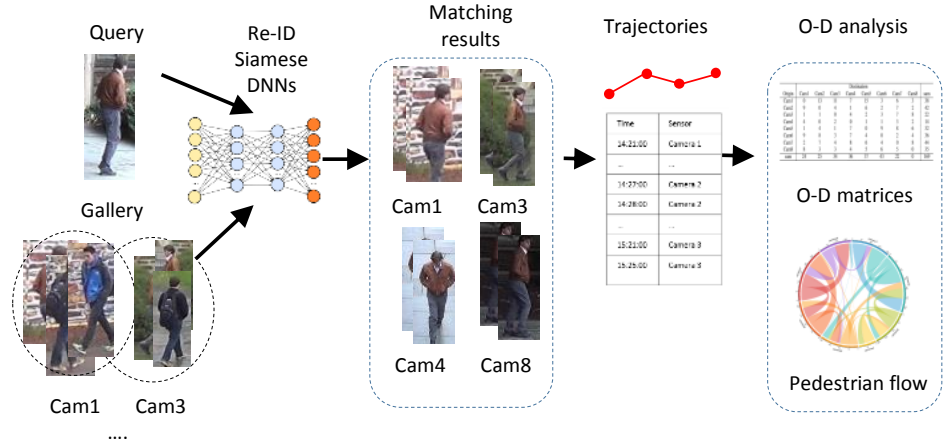


Figure 27: The proposed framework for pedestrian origin-destination estimation.

2. we validate and compare the state-of-the-art deep neural networks for pedestrian re-ID at different congestion levels.
3. we improve the accuracy in crowded scenarios by employing data augmentation techniques.

The rest of the chapter is organized as follows. Section 6.2 explains the details of the proposed framework. Section 6.3 evaluates the framework and conducted a preliminary O-D analysis based on the framework. The last section 6.4 draws conclusions and proposes future directions of this research.

6.2 METHODOLOGY

The proposed framework for pedestrian O-D estimation across multiple disjoint cameras is shown in Figure 27 and described below:

1. *Frame extraction and synchronization.* The image frames named by frame number are extracted from the video footage. Frame synchronization is to ensure that all the videos are on the same timeline so that we can associate the frames of the same time in different camera views.
2. *Detection of candidate pedestrians.* This step is a pre-process for pedestrian re-ID. In order to remove the cluttered background, the region-of-interest (ROI) of pedestrians are detected and pedestrian images are cropped from the image

frames based on the bounding boxes. All the pedestrian candidate images are collected in a *gallery*, which would be used for matching.

3. *Pedestrian re-identification (re-ID)*. The purpose of this step is to classify the detections into different identities. The approach is to initially generate the trajectories in the same camera, calculate the averaged features for each trajectory and then match the features across different cameras.
4. *Mapping the image coordinates to world coordinates and generation of pedestrian O-D flows*. The image coordinates of the pedestrian's feet are converted to world coordinates to get all the historical locations of the pedestrian. By organizing the detection in a chronological order, pedestrian O-D flows can be calculated based on the accumulated trajectories.

More details of each procedure are explained in the following sections.

6.2.1 *Frame extraction and synchronization*

First the image frames are extracted from video footage based on a certain frame interval using OpenCV packages. The frame interval is set as 120 in this paper, which is equal to 2 seconds at 60 fps frame rate. Then all the frames are named after frame count number using the native package function.

Since each camera has its own local time starting at frame 1, in order to locate the same frame in different camera views, a master camera is selected and other cameras would be synchronized to the master camera. For example 'frame_1.jpg' in camera 1 is synchronized to 'frame 5543.jpg' in camera 5. An alternative approach is that, all the frames can be named after the local time, for example 'frame_20160408_195010.jpg' means that this frame was captured at 19:50:10 on 8th of April 2016. If the recording start time of each camera and frame count number are known, the capture time of each frame can be inferred. Then it is possible to retrieve the frames at the same time stamp.

6.2.2 *Pedestrian detection and generation of gallery images*

The cluttered background in different camera views could cause confusion in the matching process. The general approach is to

detect the region-of-interests (ROI) of pedestrians from the image frames and generate the bounding box of each pedestrians. Then the pedestrian candidate images can be cropped according to the bounding boxes. All the candidate images are collected in the *gallery* for person re-ID in the next step.

The Deformable Part Model (DPM) Felzenszwalb, McAllester, and Ramanan (2008) is applied to generate the ROI automatically. The DPM model is very successful in pedestrian detection, which represents objects by component parts arranged in a deformable configuration and is less sensitive to transformations in object pose, viewpoint and nonrigid deformations.

6.2.3 Pedestrian re-identification

6.2.3.1 Data augmentation at different congestion levels

Considering practical needs of crowd management, we want to experiment with images of crowd at different levels of congestion. However a dataset with such images currently does not exist and from one dataset we cannot obtain a sufficient number of training images for different congestion levels. To overcome this issue, we enlarge the datasets by creating pedestrian images at different congestion levels, often referred to as *data augmentation*. Data augmentation techniques typically include flipping, rotating, scaling, cropping, translating, adding Guassian noise or applying Generative Adversarial Networks (GAN) to transform the images. In congested scenes, only the upper body part of the pedestrian is visible and the lower part is occluded by the surrounding people. Based on this idea, we apply image cropping and cut the lower part of the pedestrian image to simulate the effects of different congestion levels.

To determine the percentage of cropping, we approximately calculate the percentage of visible body at different congestion levels. Level of service (LOS) is the most commonly used indicator for measuring congestion. According to Fruin (1971), LOS is defined as average pedestrian space S (ft^2/p , square feet per person), as listed in the first column of Table 12. Assuming a rectangular area surrounding the pedestrian, the distance D between the pedestrians is the square root of the average pedestrian space S . Since the side of the distance (D) and the side of the visible body part (P) form a right-angled triangle, as shown in Figure 28, the percentage

of the visible body P can be calculated from the average distance D , the camera tilt angle δ and the height of the person H , as shown in Equation (12). Since the height of each pedestrian and camera tilt angle can not be obtained, in our experiments we assume an average pedestrian height of 1.70 meters and an approximate tilt angle of 45 degrees. The final calculated value of the image cropping percentage can be seen in Table 12.

With this data augmentation, for each pedestrian image, we generate six new images each representing one level of congestion which extend the original dataset and are applied to train the deep neural networks.

$$P = \frac{\tan \delta \times D}{H} \quad (12)$$



Figure 28: The idea of calculating percentage of visible body part.

LOS	Avg. space (S)	Avg. distance (D)	Visible body parts (P)
A	$> 35 \text{ ft}^2/\text{p}$	$> 1.803 \text{ m}$	whole body
B	$25 - 35 \text{ ft}^2/\text{p}$	$1.524 - 1.803 \text{ m}$	$> 90\%$ body
C	$15 - 25 \text{ ft}^2/\text{p}$	$1.180 - 1.524 \text{ m}$	$70\% - 90\%$ body
D	$10 - 15 \text{ ft}^2/\text{p}$	$0.964 - 1.180 \text{ m}$	$57\% - 70\%$ body
E	$5 - 10 \text{ ft}^2/\text{p}$	$0.682 - 0.964 \text{ m}$	$40\% - 57\%$ body
F	$< 5 \text{ ft}^2/\text{p}$	$< 0.682 \text{ m}$	$< 40\%$ body

Table 12: The relationships between LOS and the percentage of visible body parts.

6.2.3.2 Deep learning architectures for person re-ID

A siamese neural network is employed to determine if two or more pedestrian images match and can be assigned the same ID.

The network consists of three parts: the first part uses deep neural networks (DNNs) for feature extraction, then the second part performs feature tracking in single camera view and calculate the average feature of each trajectory, and finally use the loss function to match in multiple views.

In the first part, that is the robust feature extraction, we evaluated and compared three types of DNNs for feature extraction, as described in related work. We chose state-of-the-art ResNet-50 He et al. (2016), ResNet50-fc512 and ResNet50-mid Yu et al. (2017b) for global feature extraction, HA-CNN Li, Zhu, and Gong (2018) for local feature extraction and MLFN Chang, Hospedales, and Xiang (2018) for attributes based models.

Before applying the features for final matching, we perform an in-camera feature merging, which is to generate the trajectories of each ID in single camera views and then calculate the average features of each trajectory. The advantage of performing this step is that, on the one hand, the trajectory-based features have proven to be more robust than the features of a single pedestrian image, on the other hand, pedestrian re-ID in single cameras can also be used for pedestrian tracking in single camera views, which has shown good performance in MOT pedestrian tracking benchmark Zhang et al. (2017). Based on the work of Zhang et al. (2017), firstly the bounding boxes of adjacent frames are merged into tracklets by the Kuhn-Munkras algorithm in a small window. Then a hierarchical clustering is used to merge the tracklets into trajectories. Finally, averaged re-ID features of trajectories are calculated for all the trajectories in all the cameras.

In the last part, that is the matching in disjoint camera views, pairwise and triplet loss functions are compared. As displayed in Figure 29, cross-entropy is used for the pair-wise network, while triplet loss Hermans, Beyer, and Leibe (2017) is used for the triplet network.

6.2.4 *O-D flow generation*

6.2.4.1 *Coordinates transformation*

After retrieving all the matching pedestrians, the image coordinates need to be converted to world coordinates. It can be achieved by projective transformation, using at least three reference points in the images and real world maps.

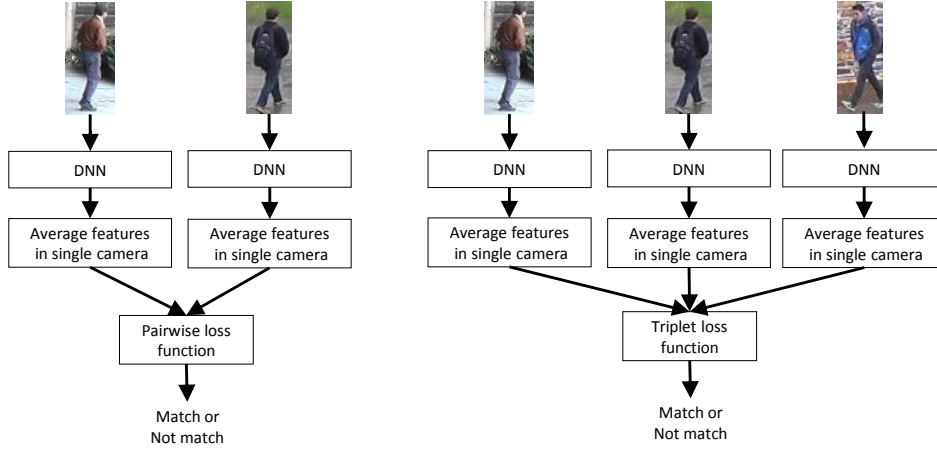


Figure 29: Pairwise siamese model takes two feature vectors as input while triplet model takes three feature vectors as inputs.

6.2.4.2 Trajectory data structure

After extracting each point, pedestrian's trajectory data is generated by connecting all the points in chronological order. In order to perform O-D analysis more efficiently, three types of points (first point, transition point and last point) are extracted and labeled, as shown in Figure 30. The point where the pedestrian was observed first is regarded as the first point. The point where the pedestrian entered or exited a new camera view, is treated as a transition point. The other point where the pedestrian was observed last, is regarded as the last point.

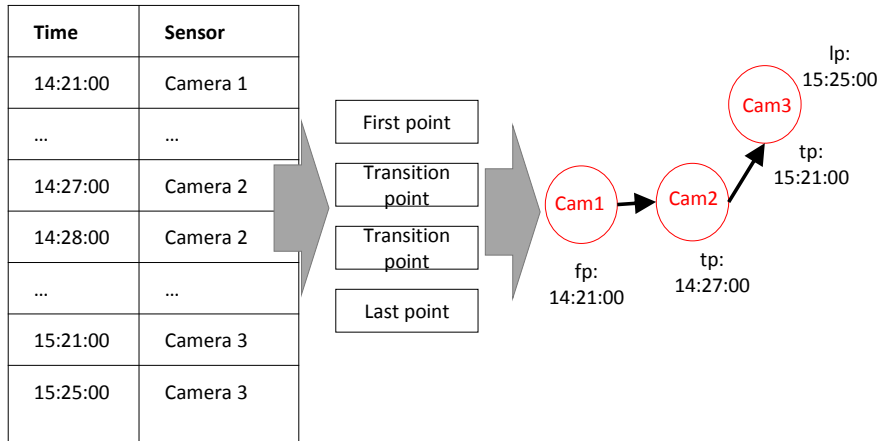


Figure 30: Trajectory data structure.

6.3 EXPERIMENTS AND RESULTS

6.3.1 Dataset

The dataset used in this study is a public dataset, DukeMTMC-reID Zheng, Zheng, and Yang (2017) and Ristani et al. (2016). It contains 85-minute high-resolution videos recorded with eight cameras on the Duke University campus, as shown in Figure 31. The location was around the Duke University Chapel, an iconic chapel in the center of the campus (the cruciform roof on the right side of Figure 31). The eight cameras covered all the sidewalks in front of the chapel to capture the pedestrian flows in and out of the chapel.

In the DukeMTMC dataset, the bounding boxes of pedestrians are provided. The dataset contains cropped pedestrian images from the videos every 120 frames, yielding in total 36,411 bounding boxes with identities. There are 1,404 identities appearing in more than two cameras and 408 identities (distractor ID) who appear in only one camera. The DukeMTMC dataset has provided the projective matrices, so we apply the matrices directly to transform the image coordinates of the pedestrian's feet to world coordinates.

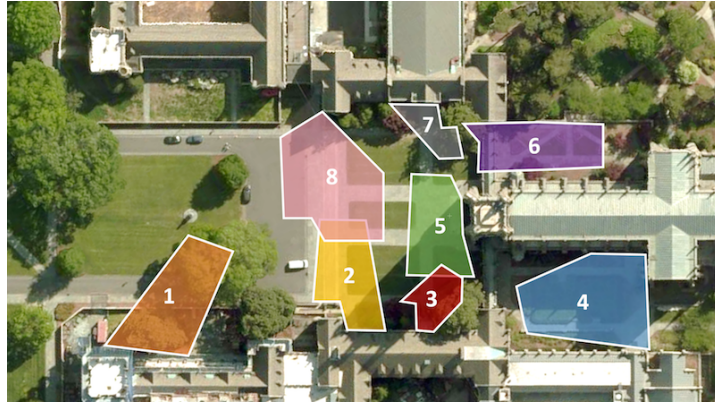


Figure 31: The camera map of DukeMTMC-reID dataset

6.3.2 Evaluation

6.3.2.1 Training strategies

702 identities were chosen as the training set and the remaining 702 identities as the testing set. In the testing set, one *query* image for each ID in each camera was selected and the remaining images

were put in the *gallery*. The training parameters used to train all the models include image height and width (256×128), optimizer (AMSGrad), learning rate (0.0003), epoch (60), step size (20 40), train batch size (32) and test batch size (100). The input image size and epoch for HA-CNN are different, which are 160×64 and 300 epochs respectively.

6.3.2.2 Model evaluation

In order to evaluate the models, we fine-tuned the models following the standard procedures for all models and reported the top-k ranking rate. The top-k ranking rate is equal to the correct matching rate that the right match appears in the top-k predictions. Cumulative Matching Characteristic (CMC) curve is a precision curve that is typically used to provide detection or recognition precision for each rank, from top-1 to top-20, as shown in Figure 32.

When evaluating which features are more robust, we found that when the congestion level is low and pedestrians are unoccluded (maximum 10% occluded), for example at LOS A and B, all models show similar performance and the five curves are close, as shown in Figure 32. But when the congestion level becomes higher and pedestrians are partially occluded (10-30% area occluded), for example LOS C and D, the performance between different models begins to differ. Overall, models based on global features, such as ResNet50, ResNet-fc512, ResNet50-mid, show better performance than local features and attributes based models at high LOS (from LOS C to F). This is despite the fact that the local features based model HA-CNN Li, Zhu, and Gong (2018) and attributes based model MLFN Chang, Hospedales, and Xiang (2018) are specifically designed models for pedestrian re-ID. So improving the performance of these models in congested scenarios should be one possible direction for future improvements.

The evaluation of loss functions shows that, the performance of ResNet50 He et al. (2016) is unfavorably affected by the triplet loss function. The ResNet-50 model with cross entropy performs better than the model with triplet loss, as shown in Table 13 and 14. Conversely, other models with triplet loss show slightly higher precision than the models with cross entropy.

Categories	Methods	LOS A	LOS B	LOS C	LOS D	LOS E	LOS F
Global	ResNet50	78.3	71.2	50.9	39.9	24.0	5.9
	ResNet-fc512	81.0	75.1	61.4	44.6	10.4	0.5
	ResNet50-mid	81.6	74.1	59.6	44.0	19.1	1.3
Local	HA-CNN	80.1	72.2	39.0	20.5	9.7	3.6
Attributes	MLFN	81.1	69.7	36.0	14.9	5.4	1.3

Table 13: Model performance using cross entropy (top-1 rate).

Categories	Methods	LOS A	LOS B	LOS C	LOS D	LOS E	LOS F
Global	ResNet50	77.7	67.8	36.6	19.9	9.6	3.2
	ResNet-fc512	80.5	75.9	64.0	48.2	21.8	1.9
	ResNet50-mid	81.5	75.0	59.0	38.3	20.6	2.4
Local	HA-CNN	79.7	72.2	39.0	20.5	9.7	3.6
Attributes	MLFN	80.4	69.4	37.1	20.1	8.6	2.4

Table 14: Model performance using triplet loss (top-1 rate).

6.3.3 O-D flow estimation

Using the transition points, we can determine where the pedestrians come from and where they go. In order to generate an O-D table, the trajectories between origin and destination are aggregated. Table 15 is an O-D table generated using extracted transition points. We can observe that the total inflow and outflow in the surveillance area of camera 2 are relatively high, exceeding 600 trips, while those in camera 4 are relatively low, around 160 trips. The flow between camera 1 and camera 2 is much higher than that between other cameras, more than 300. From the camera map, we can infer that the road connecting these two camera views is the most frequently used around the chapel.

Dynamic O-D flows over different time periods can be visualized using migration flow graphs, as shown in Figure 33. The origins and destinations of pedestrians are represented by the circle's segments. The size of the estimated flow is indicated by the width of the link as its bases and can be read using the tick marks on the outside of the circle's segments. The direction of the flow is encoded both by the origin color and by the gap between link and circle segment at the destination. We can see that the flow from camera 1 to camera 2 increased in the second half hour as

Origin	Destination								Sum
	Cam1	Cam2	Cam3	Cam4	Cam5	Cam6	Cam7	Cam8	
Cam1	0	336	0	0	12	0	4	24	376
Cam2	310	0	166	0	144	0	2	15	637
Cam3	0	158	0	160	38	1	2	0	359
Cam4	0	1	147	0	7	0	0	0	155
Cam5	20	151	45	6	0	148	23	9	402
Cam6	0	0	0	0	161	0	174	0	335
Cam7	4	0	0	0	38	248	0	155	445
Cam8	113	9	0	0	10	0	213	0	345
Sum	443	655	358	166	410	396	418	194	3040

Table 15: Origin-destination (O-D) table

indicated by the larger width of the blue arc. From the flow scale around the circle, we can also observe that the total flow within the coverage of camera 2 was the largest as indicated by the longest arc length. Furthermore, since the red arc length became shorter, the total pedestrian flow in the coverage area of the camera 7 reduced in the second half hour as indicated by the shorter length of the red arc.

6.4 CONCLUSIONS AND FUTURE PERSPECTIVES

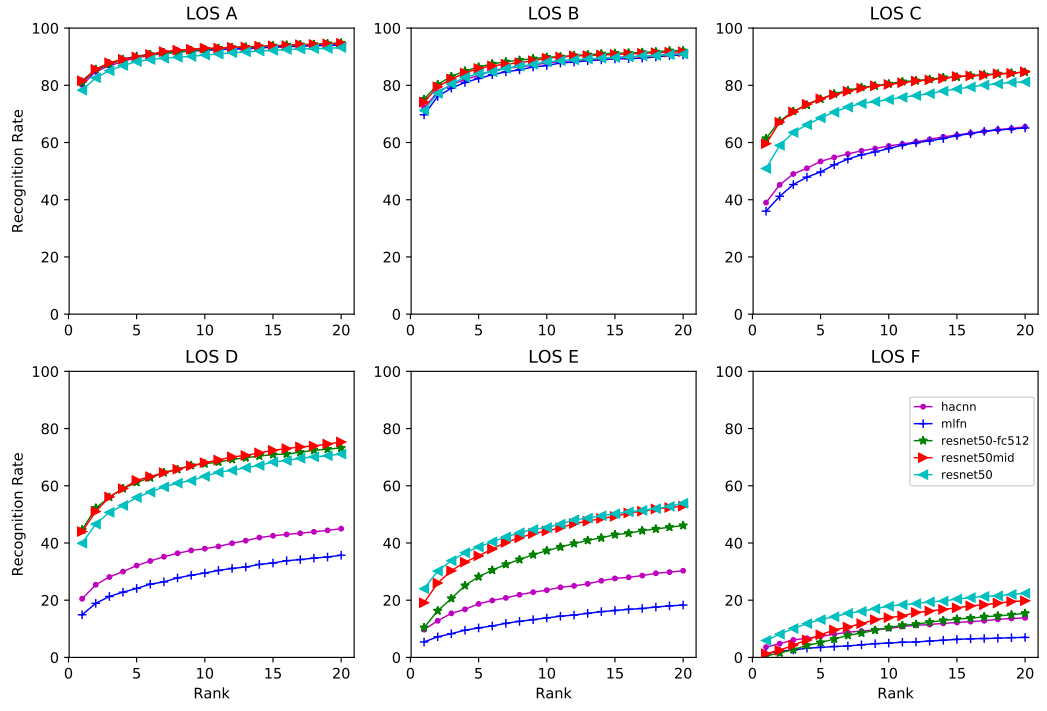
In this study, we proposed a novel pedestrian O-D estimation framework based on multi-camera pedestrian re-identification (re-ID). Combined with the applications in transport infrastructure management, we evaluated and compared the state-of-the-art models of pedestrian re-ID at different congestion levels. Data augmentation technique was applied to improve the accuracy in crowded scenarios. We believe this work can provide new insights into automated pedestrian flow estimation and stimulate the research on intelligent transport infrastructure management.

In the future work, this research would be continued and extended in several ways:

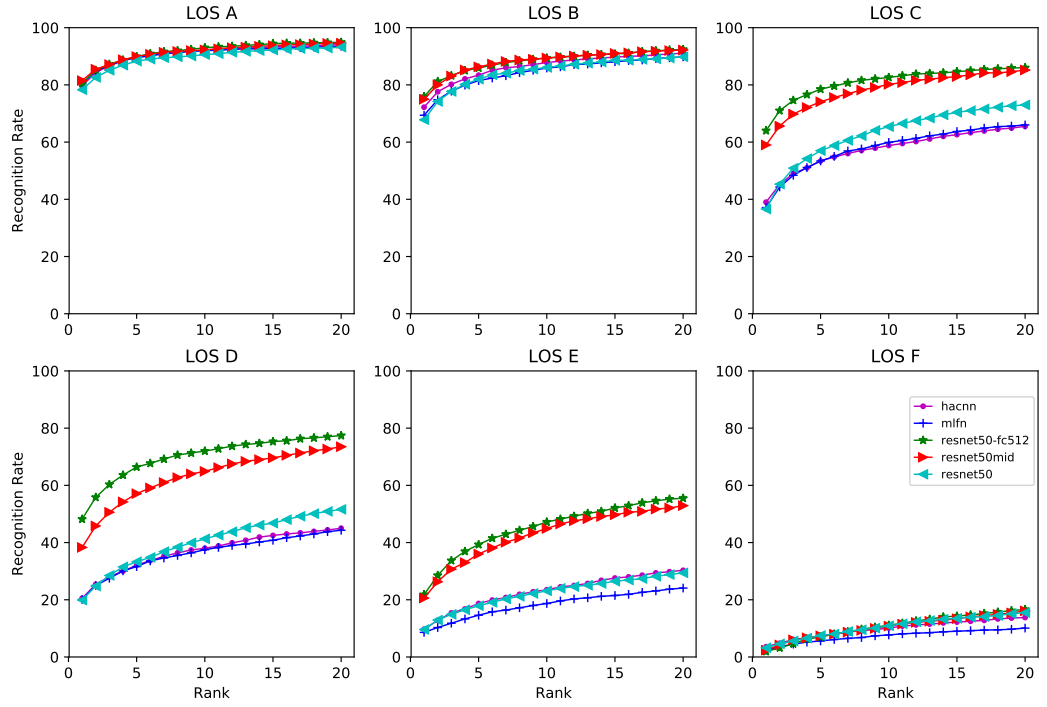
- The entire network for multi-view pedestrian re-ID can be designed in an end-to-end fashion. Since the pedestrian detection and the pedestrian re-ID were processed separately

in this article, they may share the same backbone for feature extraction, which can make the framework more efficient when processed online. The extracted features can be used not only to generate the bounding box of a pedestrian but also for subsequent matching.

- An additional input channel can be applied, that is the depth information collected by the depth sensors, and integrate the information into the current structure. The depth image contains information relating to the distance of the surfaces of pedestrians from a viewpoint, from which skeleton information and anthropometric measurements of the pedestrians can be extracted, which have been proven to be more robust than appearance features extracted from RGB cameras in different camera environments.
- State-of-the-art pedestrian re-ID techniques take the bounding boxes of pedestrians as inputs. Since the background is useless in the pedestrian re-id process, image segmentation can be combined with pedestrian detection to generate a more accurate pedestrian shape and completely remove the background.



(a) Cross-entropy loss



(b) Triplet loss

Figure 32: CMC curve of image-based methods under different congestion levels.

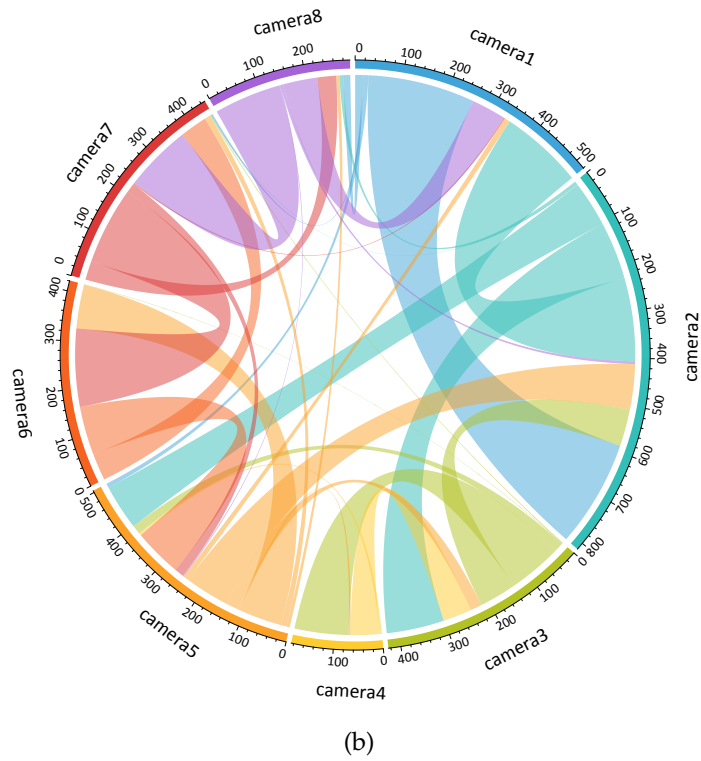
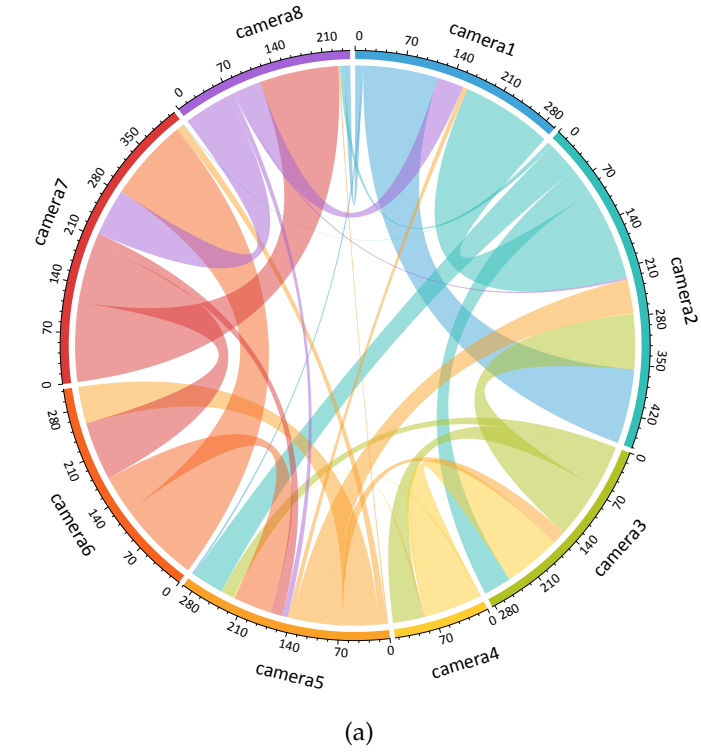


Figure 33: Pedestrian O-D flow in the first half hour (a) and the second half (b) of the video.

PEDESTRIAN ORIGIN-DESTINATION ANALYSIS USING DEPTH CAMERAS

This chapter is based on the paper entitled "Pedestrian origin-destination estimation in emergency scenarios" published in *2019 9th International Conference on Fire Science and Fire Protection Engineering (ICFSFPE)*, which received the *First Prize (Excellent Paper Award)*. I was conductive at every stage of this work, including research concept proposal, experiment design and implementation, data analysis, discussions and paper writing. My supervisor Prof. Majid Sarvi and Dr. Kourosh Khoshelham was providing a sufficient amount of support in discussing research ideas, supervising the progress, and discussing the results.

7.1 INTRODUCTION

The unpredictable pedestrian flow in an emergency can lead to accidents and even fatalities. Extracting the information of the volume of people travelling from a group of origin zones to a group of destination zones, known as Origin-Destination (O-D) information, is pivotal in emergency management. Because of challenges regarding poor lighting conditions in fires, and occlusion in dense scenarios, recent research has focused on top view and RGB-D based camera solutions. In this paper, we propose an innovative RGB-D based pedestrian O-D estimation framework. The main contributions of this study is three-fold:

1. A joint convolutional neural network (CNN) is trained on color, depth and infrared images to generate a set of multi-modal fused features for pedestrian re-identification.
2. For evaluation we collect a new RGB-D dataset considering different levels of service and lighting conditions.
3. The proposed method achieves a rank-1 accuracy of 80.5% under good lighting conditions, which is a 13.4% improvement as compared to using a single RGB modality. Even in low lighting conditions where RGB cameras perform poorly,

we can still achieve an accuracy of 72.8% using the depth cameras.

The remaining part of the chapter is organized as follows. Section 7.2 introduced the multi-modality framework for origin-destination estimation. Section 7.3 described the dataset we collected for the experiment. The evaluation on the dataset and findings of the experiment is explained in Section 7.4. The conclusions and future work are discussed in Section 7.5.

7.2 METHODOLOGY

In order to integrate color, depth and infrared information and take the advantage of CNNs, we propose a framework that can jointly process the three modalities, RGB, depth and infrared simultaneously. Based on Eitel et al., 2015, we apply an architecture that has three parallel CNN streams processing three input images separately while being concatenated in a later fully connected layer, as shown in Figure 34. Each CNN stream uses the AlexNet architecture and consists of five convolution layers and three fully connected layers following the last convolution layer and transforming the feature maps to sparse representations for RGB, depth and infrared modalities separately. The feature representations of three modalities (fc7RGB, fc7D and fc7I) are fused and used as input to a fully connected layer (fc8) which learns a joint RGB-D-I feature representations from colour, depth and infrared images. Lastly, the output probabilities for each class, defined as a person identity are calculated by a softmax layer (fc9). A loss function is used to update the parameters of the network. When testing, we remove the softmax layer and extract features from fc8.

Firstly, three identical CNN models are trained for RGB, depth and infrared modalities before training the whole architecture, referred as RGB-CNN, Depth-CNN and Infrared-CNN. A pre-trained model of AlexNet trained on the ImageNet dataset is used to initialize the model weights. Following the architecture of AlexNet, the size of input image is 227×227 . Then all color, depth and infrared images are adjusted accordingly before being processed by the network.

Furthermore, the model parameters of RGB-CNN, Depth-CNN and Infrared-CNN are used to initialize the three CNN streams in the architecture. The softmax layers are replaced by a randomly

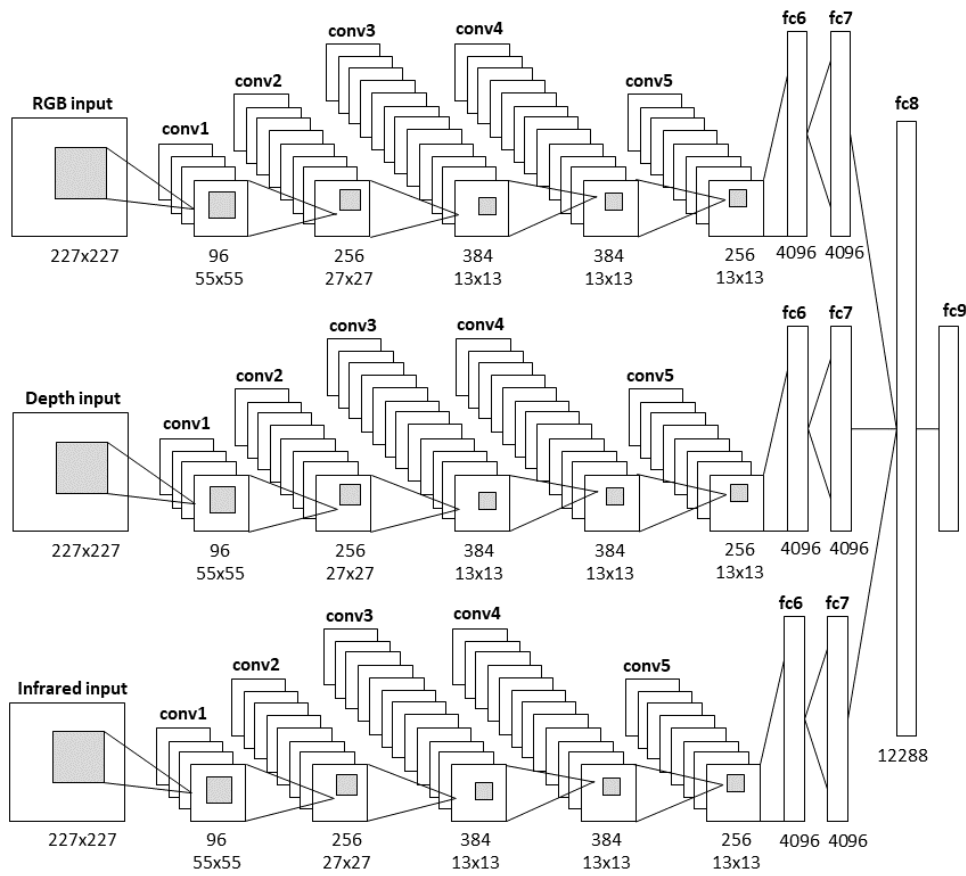


Figure 34: Overview of the framework. The color, depth and infrared images are processed separately and fused in the last fully connected layer to generate a fused feature for classification.

initialized fully connected layer (fc8) and new softmax layer (fc9). By fusing outputs from fc7RGB, fc7D and fc7I in fc8, the parameters of one stream are updated depending on the input to other two streams, while the weights and bias of fc8 are updated based on the three inputs, resulting in a fused output.

7.3 DATASETS

To evaluate the proposed framework, we present a novel RGB-D dataset from a top view, called Crowd RGB-D dataset.

7.3.1 Camera setup

Four sets of RGB-D cameras are installed in the experiment, including a ZED stereo camera from Stereolabs and three Microsoft Kinect V2 Sensors. Microsoft Kinect belongs to Time of Flight (ToF) camera, which obtains the target distance by transmitting a light pulse to the target, then receive the light returned from the object with the sensor, and detect the flight (round trip) time of the light pulse. The ZED camera is a stereo camera that imitates the structure of the human eye and determines the distance information by the parallax of the two cameras. The measuring range of ZED camera is 0.7m-20m which can cover both low and high ceilings, while the Kinect sensor can record depth from a range 0.5m-4.5m.

The ZED camera and one Kinect sensor were installed at the classroom doorway in order to compare the performances of the two types of depth cameras. Two Kinect cameras were installed in the two evacuation exits separately. The location map is shown in Figure 35.

7.3.2 Scenarios

Assuming that the fire alarms in the building are turned on, 16 identities are invited to evacuate from the classroom to two exits at six levels of service under both good and poor lighting conditions. As a result, cases from low to high crowd densities can be recorded, snapshots of captured color images, depth images and infrared images in good and poor lighting conditions are shown in Table 16 and 17.

Table 16: Snapshots at classroom doorway (poor lighting). The color cameras perform very poorly in this situation.

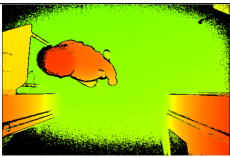
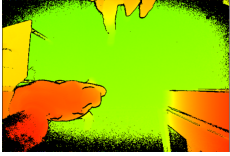
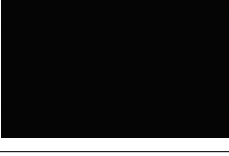
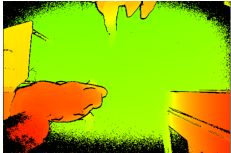
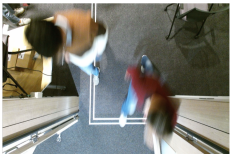
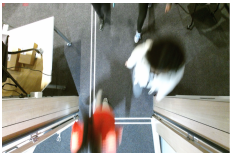
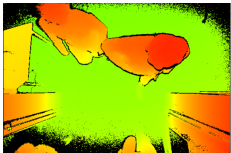
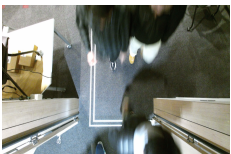
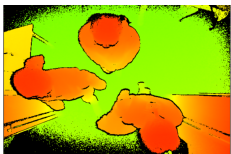
LOS	Color image	Depth image	Infrared image
A			
B			
C			
D			
E			
F			

Table 17: Snapshots at classroom doorway (good lighting)

LOS	Color image	Depth image	Infrared image
A			
B			
C			
D			
E			
F			

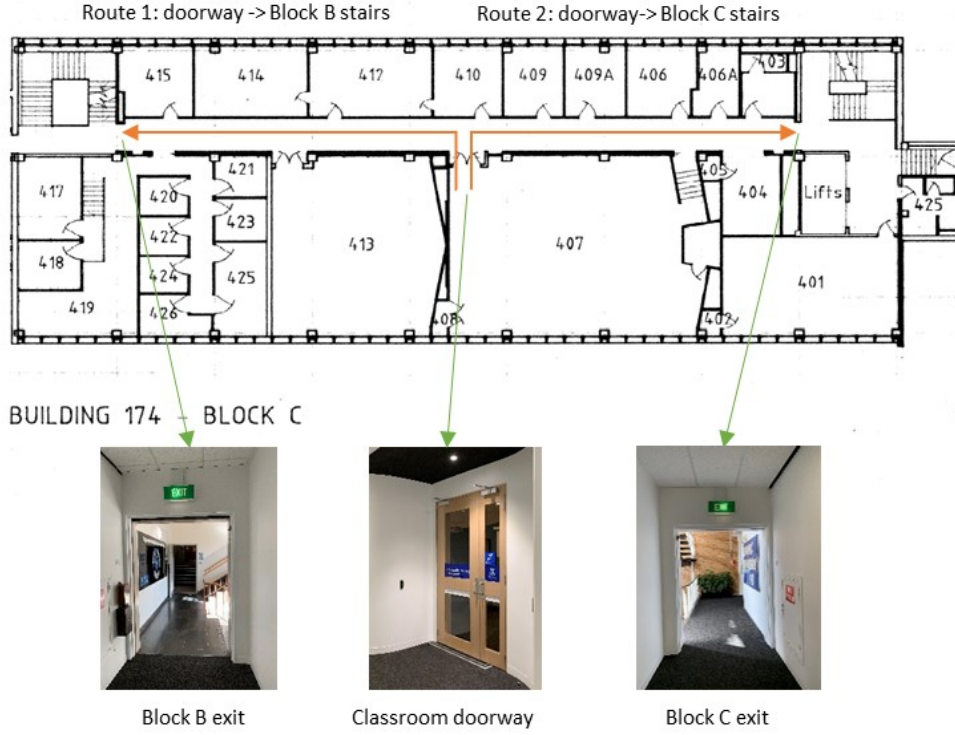


Figure 35: Evacuation map

7.3.3 Ground-truths

We manually annotate the bounding box of each person and use these annotations to evaluate the person re-ID algorithm and O-D estimation framework.

7.4 EXPERIMENTS AND RESULTS

8 identities were chosen as the training set and the remaining 8 identities as the testing set. After training the CNN models, features from the test set are extracted from the last fully connected layer. When RGB-CNN, Depth-CNN and Infrared-CNN are trained individually, we use a batch size of 128 while a size of 64 for RGB-D-I-CNN. Stochastic Gradient Descent (SGD) with momentum is used to update the network parameters and avoid getting stuck in a local minimum. Hyper parameters are set accordingly to Similar to Eitel et al., 2015, a momentum of 0.9 and learning rate of 0.01 which is reduced by multiplying with 0.97 for each epoch are set. The person re-identification results are presented by calculating the rank-1 accuracy based on feature matching where rank-1 indicates the correct matching rate that the right

match is the top-1 prediction. Then the correct match makes a right O-D pair.

Table 18 and 19 show the rank-1 accuracy of pedestrian re-ID in both good and poor lighting conditions. It is evident that fusing RGB, depth and infrared modalities improves accuracy compared to using a single modality. The RGB-D-I-CNN method improve the accuracy by 13.4% compared with the RGB-CNN method. For single modality, depth modality shows the lowest performance, while infrared-CNN shows better performance than RGB-CNN method. We can see from Table 17 that when the LOS is low, the walking speed of pedestrians increases and the color images become blurry. However, the depth images are clear, which could result in higher recognition rate.

Figure 36 shows the results of O-D detections. At different levels of service, the number of correct O-D detections is not much different. The reason is that from the top view, the upper body of each identity can be recognized clearly so the occlusion issue caused by the dense scenarios does not exist.

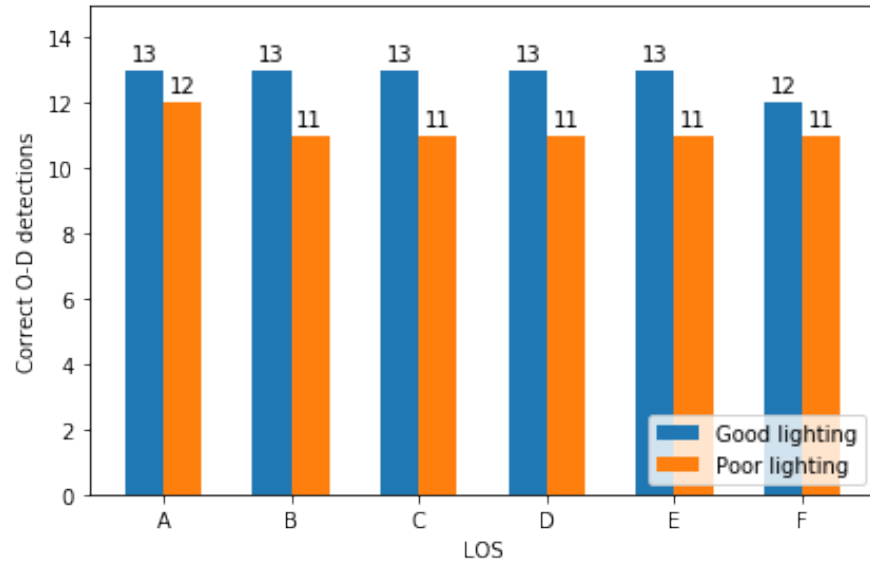


Figure 36: Correct O-D detections by LOS in good and poor lighting conditions based on RGB-D-I-CNN method. There are 16 O-D routes per LOS.

In comparison of different lighting conditions, we can see in Table 16 that the color cameras perform very poorly in low lighting conditions, so the features extracted from color images are all missing, which affect the results of both RGB-CNN method and fusion RGB-D-I-CNN method.

Table 18: Rank-1 accuracy in good lighting condition (%)

LOS	RGB-CNN	Depth-CNN	Infrared-CNN	RGB-D-I-CNN
A	68.1	25.1	74.1	80.5
B	66.2	22.3	73.1	79.6
C	66.8	24.1	73.6	78.9
D	67.2	22.6	74.5	78.6
E	66.5	23.9	73.8	78.7
F	66.9	22.8	74.2	77.4

Table 19: Rank-1 accuracy in poor lighting condition (%)

LOS	RGB-CNN	Depth-CNN	Infrared-CNN	RGB-D-I-CNN
A	-	24.8	64.2	72.8
B	-	23.9	63.4	70.5
C	-	24.7	62.8	70.2
D	-	22.9	63.9	71.3
E	-	21.6	64.7	71.2
F	-	22.4	61.6	69.6

7.5 CONCLUSIONS AND FUTURE WORK

In this study, we proposed a RGB-D based CNN for pedestrian O-D estimation in emergency scenarios. Three CNN models are trained on colour, depth and infrared images, respectively, and a joint RGB-D-I-CNN model is initialized using the resulting trained parameters. To evaluate the framework, we collected a novel top view RGB-D dataset at different levels of service in both good and poor lighting conditions. By applying our framework to the new dataset, it is evident that the fusion of RGB, depth and infrared modalities can improve the accuracy by 13.4% compared with using single RGB modality. Furthermore, when color cameras perform very poorly in low lighting conditions, we can still achieve an accuracy of 72.8% , which proves the effectiveness of RGB-D cameras in emergency scenarios, especially fire and low visibility scenes.

For the future work, the research would continue in three directions:

- to discover more possibilities from depth images by extracting anthropometric features, which are unique among people.
- to compare the performances of ZED camera and Kinect sensor.
- to understand the pedestrian evacuation choices.

Part IV

CONCLUSION AND FUTURE WORK

Part [iv](#) concludes the thesis. A general conclusion of the thesis along with the major contributions is drawn in Chapter [8](#). Chapter [9](#) presents high-level discussions based on the experimental results, and recommendation for future research into the future the potential trend of visual congestion analysis and technical challenges.

CONCLUSION

The thesis contributes to the knowledge of crowd congestion management in both macro-scale and micro-scale by exploring the visual congestion analysis techniques and evaluating the proposed models. Experiment results suggest that the proposed multi-scale, multi-view and multi-modality models contribute to more accurate, efficient and intelligent station management.

8.1 RESEARCH CONTRIBUTIONS

In summary, the main contributions of the research presented in this thesis can be classified in the following broad categories:

1. An operationalization of the crowd modeling theory in the station management and image processing domain. This interpretation is based on the understanding and defining the strengths and weaknesses of the current visual congestion analysis in the context of smart station management.
2. Spatial-temporal congestion models, presenting an image processing perspective to the detection and prediction of the crowd congestion integrating both spatial and temporal dependencies.
3. The development of the framework of congestion management that can potentially be used to support the design and development of industry road maps and associated strategies to facilitate the realization of vision-based congestion management systems.

8.2 MAJOR RESULTS AND FINDINGS

8.2.1 *Macro-scale congestion*

Understanding congestion at a macro scale provides a big picture of congestion distribution and location, so that further crowd control and management can be implemented. The pedestrian congestion map is the most popular method to represent pedestrian flow.

This thesis starts from the crowd congestion map generation in single camera view, further extended to the model of multi-view map fusion. Finally a map-based prediction model is proposed.

In single camera view, a novel approach to direct classification of level of service using CCTV images is proposed, implemented and tested with applications in congestion management of major transport stations. The direct one-step methods avoid complex calculation of detailed flow parameters and greatly improve the detection efficiency therefore achieve real-time monitoring. Results of an experiment in Flinders Street Station, Melbourne, Australia show that congestion levels can be accurately identified with a total accuracy of 81.9%. It was also shown that exploiting the temporal correlation between successive frames results in more plausible level of service maps with fewer temporal fluctuations. The proposed approach achieves near real-time performance with an average computation time of 0.40 seconds per frame. Our results demonstrate that LOS maps updated in real time provide invaluable information for monitoring crowd congestion in transport stations and improving the safety of passengers.

A multi-view fusion framework of generating crowd congestion maps is proposed to provide insights into macroscopic crowd management in large public space. The multi-view fusion methods based on ensemble rules were compared and evaluated. The results show that algebraic methods outperform voting methods and borda count, because algebraic methods consider both the probabilities and the results of all camera views. Two datasets both in sparse and dense scenarios were tested. The final results show that the fusion of camera views can not only provide a wide coverage of monitored space but also help improve the classification accuracy of individual views, especially the occluded views. A loading test was conducted to evaluate the maximum number of cameras on our test environment. With a map refresh rate of 2 seconds, 48 cameras can be processed at the same time, which means that a density map covering 48 cameras views would be visualized.

We applied deep learning method to the challenging problem of short-term congestion prediction, which so far has not been studied from an image analysis perspective based on open source congestion maps. We formulate congestion prediction as a spatio-temporal sequence forecasting problem and introduce ConvLSTM structure to tackle the problem. The advantage of our approach is

that unlike the previous methods that involve the laborious work of fusing many traffic parameters, our method requires only historical congestion maps for congestion analysis. Based on our experiment on the freeway and urban minor roads, this framework can predict 15 minutes ahead on the highway with an accuracy of 88.1% and 1 hour ahead on the urban minor roads with an accuracy of 78.3%.

8.2.2 *Micro-scale congestion*

At a micro scale, the trajectories of individuals are extracted and analysed. We proposed a novel pedestrian origin-destination (O-D) estimation framework based on multi-camera pedestrian re-identification (re-ID). By re-identifying the same person in different camera views, the O-D of each identity can be generated. Combined with the applications in transport infrastructure management, we evaluated and compared the state-of-the-art models of pedestrian re-ID at different congestion levels. Data augmentation technique was applied to improve the accuracy in crowded scenarios. We believe this work can provide new insights into automated pedestrian flow estimation and stimulate the research on intelligent transport infrastructure management.

To explore alternative visual sensors for congestion analysis, a RGB-D based CNN for pedestrian O-D estimation in emergency scenarios is proposed. Three CNN models are trained on colour, depth and infrared images, respectively, and a joint RGB-D-I-CNN model is initialized using the resulting trained parameters. To evaluate the framework, we collected a novel top view RGB-D dataset at different levels of service in both good and poor lighting conditions. By applying our framework to the new dataset, it is evident that the fusion of RGB, depth and infrared modalities can improve the accuracy by 13.4% compared with using single RGB modality. Furthermore, when color cameras perform very poorly in low lighting conditions, we can still achieve an accuracy of 72.8% , which proves the effectiveness of RGB-D cameras in emergency scenarios, especially fire and low visibility scenes.

LIMITATIONS AND FUTURE WORK

9.1 INTERNET OF VISUAL SENSORS

As the monitored space becomes complex and the size of camera network fast increases, there are higher requirements for the robustness, efficiency, flexibility of the multi-view fusion schema. Our experiments demonstrated that the overall accuracy of crowd congestion map generation can be improved by the multi-view fusion framework. More factors can be considered to improve the accuracy, e.g. camera angle. With higher camera angle, the occlusion can be reduced and more people can be seen in the camera view. An efficiency test is also conducted in this paper to test the feasibility of the application. The percentage of overlapping camera views is an important factor in the efficiency test as larger overlapping areas require longer fusion time. It can be seen that the a good design of the internet of visual sensors is indispensable for large area crowd congestion management as it can not only improve fusion accuracy but also efficiency.

9.2 FROM TRAFFIC TO CROWD CONGESTION PREDICTION

The presented results on traffic congestion prediction are promising and this research can stimulate the research on traffic prediction from an image processing perspective, which is very useful for understanding the spatio-temporal dependencies based on realistic road networks. Due to the unavailability of large training dataset, for example, congestion maps for one year, we test the prediction model on traffic congestion maps first. For the future work, the collection of crowd congestion maps would continue so that the model test on the crowd congestion can be performed and the crowd congestion patterns can be analyzed. Meanwhile, the prediction model can be improved in several directions:

- The model can be extended to more congestion parameters such as pedestrian speed and flow.

- In order to learn the spatial-temporal dependencies in a wider area, with more computing resources, the depth of the network and the input image size can be increased which can greatly improve the performance of the model.
- It is difficult for us to compare with traditional non-visual methods, because there is no pedestrian flow data other than CCTV videos. So we only compare with visual methods, such as different image processing techniques. If we can obtain more data, we would carry out this work in the future.

9.3 MULTI-MODALITY ORIGIN-DESTINATION ESTIMATION

The research on pedestrian origin-destination estimation using only CCTV cameras can be continued and extended in several ways:

- The entire network for multi-view pedestrian re-ID can be designed in an end-to-end fashion. Since the pedestrian detection and the pedestrian re-ID were processed separately in this article, they may share the same backbone for feature extraction, which can make the framework more efficient when processed online. The extracted features can be used not only to generate the bounding box of a pedestrian but also for subsequent matching.
- State-of-the-art pedestrian re-ID techniques take the bounding boxes of pedestrians as inputs. Since the background is useless in the pedestrian re-id process, image segmentation can be combined with pedestrian detection to generate a more accurate pedestrian shape and completely remove the background.

Additional input channels can provide more information about pedestrians, for example, depth information collected by the RGB-D sensors, which can be integrated into the RGB only structure. The depth image contains information relating to the distance of the surfaces of pedestrians from a viewpoint, from which skeleton information and anthropometric measurements of the pedestrians can be extracted, which have been proven to be more robust than appearance features extracted from CCTV cameras in different camera environments. Towards a multi-modality framework, the research on RGB-D based pedestrian origin-destination estimation in this study can continue in three directions:

- to discover more possibilities from depth images, for example, by extracting anthropometric features, which are unique among people.
- to compare the performances of ZED camera and Kinect camera installed in the experiment.
- to understand the pedestrian evacuation choices.

To conclude, the research presented in this thesis demonstrated the great potential of visual crowd analysis for improving the safety of the crowd and enhancing the management of transport hubs.

BIBLIOGRAPHY

- Alahi, Alexandre, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese (2016). "Social lstm: Human trajectory prediction in crowded spaces." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971 (cit. on p. 18).
- Amiri, S.M., M.T. Pourazad, P. Nasiopoulos, and V.C.M. Leung (2013). "Non-intrusive human activity monitoring in a smart home environment." In: *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services, Healthcom 2013* (cit. on p. 27).
- Asano, Miho, Agachai Sumalee, Masao Kuwahara, and Shinji Tanaka (2007). "Dynamic cell transmission-based pedestrian model with multidirectional flows and strategic route choices." In: *Transportation Research Record: Journal of the Transportation Research Board* 2039, pp. 42–49 (cit. on p. 33).
- Atrey, Pradeep K, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli (2010). "Multimodal fusion for multimedia analysis: a survey." In: *Multimedia systems* 16.6, pp. 345–379 (cit. on p. 15).
- Azadbakht, Mohsen, Clive S Fraser, and Kourosh Khoshelham (2018). "Synergy of sampling techniques and ensemble classifiers for classification of urban environments using full-waveform LiDAR data." In: *International Journal of Applied Earth Observation and Geoinformation* 73, pp. 277–291 (cit. on p. 15).
- Bai, Xiang, Mingkun Yang, Tengteng Huang, Zhiyong Dou, Rui Yu, and Yongchao Xu (2017). "Deep-Person: Learning Discriminative Deep Features for Person Re-Identification." In: *arXiv preprint arXiv:1711.10658* (cit. on p. 20).
- Balderas, David, Pedro Ponce, and Arturo Molina (2019). "Convolutional long short term memory deep neural networks for image sequence prediction." In: *Expert Systems with Applications* 122, pp. 152–162 (cit. on p. 18).
- Barbosa, Igor Barros, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino (2012). "Re-identification with rgb-d sensors." In: *European Conference on Computer Vision*. Springer, pp. 433–442 (cit. on pp. 22, 26).

- Batchuluun, Ganbayer, Rizwan Ali Naqvi, Wan Kim, and Kang Ryoung Park (2018). "Body-movement-based human identification using convolutional neural network." In: *Expert Systems with Applications* 101, pp. 56–77 (cit. on p. 18).
- Boltes, Maik and Armin Seyfried (2013). "Collecting pedestrian trajectories." In: *Neurocomputing* 100, pp. 127–133 (cit. on p. 18).
- Breiman, Leo (1996). "Bagging predictors." In: *Machine learning* 24.2, pp. 123–140 (cit. on p. 16).
- Candamo, Joshua, Matthew Shreve, Dmitry B. Goldgof, Deborah B. Sapper, and Rangachar Kasturi (2010). "Understanding transit scenes: A survey on human behavior-recognition algorithms." In: *IEEE Transactions on Intelligent Transportation Systems* 11.1, pp. 206–224 (cit. on p. 23).
- Chang, Xiaobin, Timothy M Hospedales, and Tao Xiang (2018). "Multi-level factorisation net for person re-identification." In: *CVPR*. Vol. 1, p. 2 (cit. on pp. 21, 82, 85).
- Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz (2015). "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor." In: *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, pp. 168–172 (cit. on p. 27).
- Cheng, Zhongwei, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian (2012). "Human daily action analysis with multi-view and color-depth data." In: *European Conference on Computer Vision*. Springer, pp. 52–61 (cit. on pp. 23, 27).
- Chollet, François et al. (2015). *Keras*. <https://github.com/fchollet/keras> (cit. on p. 68).
- Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks." In: *Machine learning* 20.3, pp. 273–297 (cit. on p. 15).
- Davies, Anthony C, Jia Hong Yin, and Sergio A Velastin (1995). "Crowd monitoring using image processing." In: *Electronics & Communication Engineering Journal* 7.1, pp. 37–47 (cit. on p. 13).
- Eitel, Andreas, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard (2015). "Multimodal deep learning for robust RGB-D object recognition." In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 681–687 (cit. on pp. 92, 97).
- Ermagun, Alireza and David Levinson (2018). "Spatiotemporal traffic forecasting: review and proposed directions." In: *Transport Reviews* 38.6, pp. 786–814 (cit. on p. 70).

- Feliciani, Claudio and Katsuhiro Nishinari (2018). "Measurement of congestion and intrinsic risk in pedestrian crowds." In: *Transportation Research Part C Emerging Technologies* 91.JUN. Pp. 124–155 (cit. on p. 13).
- Felzenszwalb, Pedro, David McAllester, and Deva Ramanan (2008). "A discriminatively trained, multiscale, deformable part model." In: (cit. on p. 80).
- Ferryman, J and A Shahrokni (2009). "Pets2009: Dataset and challenge." In: *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*. IEEE, pp. 1–6 (cit. on p. 51).
- Firman, Michael (2016). "RGBD datasets: Past, present and future." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 19–31 (cit. on p. 21).
- Freund, Yoav, Robert Schapire, and Naoki Abe (1999). "A short introduction to boosting." In: *Journal-Japanese Society For Artificial Intelligence* 14.771-780, p. 1612 (cit. on p. 16).
- Fruin, J.J. and G.R. Strakosch (1987). *Pedestrian Planning and Design*. Elevator World (cit. on pp. 11, 12).
- Fruin, John J (1971). *Pedestrian planning and design*. Tech. rep. (cit. on p. 80).
- Gardziński, Paweł, Krzysztof Kowalak, Sławomir Maćkowiak, and Łukasz Kamiński (2015). "Crowd density estimation based on voxel model in multi-view surveillance systems." In: *Systems, Signals and Image Processing (IWSSIP), 2015 International Conference on*. IEEE, pp. 216–219 (cit. on p. 15).
- Gayathri, H, PM Aparna, and Ashish Verma (2017). "A review of studies on understanding crowd dynamics in the context of crowd safety in mass religious gatherings." In: *International journal of disaster risk reduction* 25, pp. 82–91 (cit. on p. 4).
- Grant, Jason M and Patrick J Flynn (2017). "Crowd scene understanding from video: a survey." In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13.2, p. 19 (cit. on p. 4).
- Guo, Renyong, Haijun Huang, and SC Wong (2011). "Collection, spillback, and dissipation in pedestrian evacuation: A network-based method." In: *Transportation research part B: methodological* 45.3, pp. 490–506 (cit. on p. 33).
- Hall, David L and James Llinas (1997). "An introduction to multisensor data fusion." In: *Proceedings of the IEEE* 85.1, pp. 6–23 (cit. on p. 16).

- Haque, Albert, Alexandre Alahi, and Li Fei-Fei (2016). "Recurrent attention models for depth-based person identification." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1229–1238 (cit. on pp. 22, 26).
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on pp. 14, 20, 82, 85).
- He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick (2017). "Mask r-cnn." In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, pp. 2980–2988 (cit. on p. 52).
- Hermans, Alexander, Lucas Beyer, and Bastian Leibe (2017). "In defense of the triplet loss for person re-identification." In: *arXiv preprint arXiv:1703.07737* (cit. on pp. 19, 82).
- Hofmann, Martin, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll (2014). "The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits." In: *Journal of Visual Communication and Image Representation* 25.1, pp. 195–206 (cit. on pp. 22, 26).
- Huang, Gao, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger (2017). "Densely connected convolutional networks." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (cit. on p. 14).
- Ibrahim, Azhar Mohd, Ibrahim Venkat, KGB Subramanian, Ahamad Tajudin Khader, and Philippe De Wilde (2016). "Intelligent evacuation management systems: A review." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.3, p. 36 (cit. on p. 4).
- Idrees, Haroon, Imran Saleemi, Cody Seibert, and Mubarak Shah (2013). "Multi-source multi-scale counting in extremely dense crowd images." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2547–2554 (cit. on p. 52).
- Jacobs, Robert A, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton (1991). "Adaptive mixtures of local experts." In: *Neural computation* 3.1, pp. 79–87 (cit. on p. 16).
- Kaiser, Mohammed Shamim, Khin T Lwin, Mufti Mahmud, Donya Hajializadeh, Tawee Chaipimonplin, Ahmed Sarhan, and Mohammed Alamgir Hossain (2017). "Advances in crowd analysis for urban applications through urban event detection." In: *IEEE Transactions on Intelligent Transportation Systems* 19.10, pp. 3092–3112 (cit. on p. 4).

- Kalayeh, Mahdi M, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah (2018). "Human Semantic Parsing for Person Re-identification." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1062–1071 (cit. on p. 20).
- Kastaniotis, Dimitris, Ilias Theodorakopoulos, Christos Theoharatos, George Economou, and Spiros Fotopoulos (2015). "A framework for gait-based recognition using Kinect." In: *Pattern Recognition Letters* 68, pp. 327–335 (cit. on pp. 22, 26).
- Ke, Jintao, Hongyu Zheng, Hai Yang, and Xiqun Michael Chen (2017). "Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach." In: *Transportation Research Part C: Emerging Technologies* 85, pp. 591–608 (cit. on p. 17).
- Koppula, Hema Swetha, Rudhir Gupta, and Ashutosh Saxena (2013). "Learning human activities and object affordances from rgb-d videos." In: *The International Journal of Robotics Research* 32.8, pp. 951–970 (cit. on p. 27).
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*, pp. 1097–1105 (cit. on pp. 14, 17).
- Lam, William HK, Chung-Yu Cheung, and CF Lam (1999). "A study of crowding effects at the Hong Kong light rail transit stations." In: *Transportation Research Part A: Policy and Practice* 33.5, pp. 401–415 (cit. on pp. 11, 12).
- Lamba, Sonu and Neeta Nain (2017). "Crowd monitoring and classification: a survey." In: *Advances in Computer and Computational Sciences*. Springer, pp. 21–31 (cit. on p. 4).
- Lejbølle, Aske R, Kamal Nasrollahi, Benjamin Krogh, and Thomas B Moeslund (2017). "Multimodal Neural Network for Overhead Person Re-identification." In: *BIOSIG 2017* (cit. on pp. 22, 26).
- Li, Jingwen, Lei Huang, and Changping Liu (2012). "People counting across multiple cameras for intelligent video surveillance." In: *Proceedings - 2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2012*, pp. 178–183 (cit. on p. 15).
- Li, Wanqing, Zhengyou Zhang, and Zicheng Liu (2010). "Action recognition based on a bag of 3d points." In: *Computer Vision*

- and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.* IEEE, pp. 9–14 (cit. on p. 27).
- Li, Wei, Xiatian Zhu, and Shaogang Gong (2018). “Harmonious attention network for person re-identification.” In: *CVPR*. Vol. 1, p. 2 (cit. on pp. 20, 82, 85).
- Li, Yaguang, Rose Yu, Cyrus Shahabi, and Yan Liu (2018a). “Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting.” In: *International Conference on Learning Representations* (cit. on p. 17).
- Li, Yan, Majid Sarvi, Kourosh Khoshelham, and Milad Haghani (2018b). *Real-Time Level-of-Service Maps Generation from CCTV Videos*. Tech. rep. (cit. on pp. 16, 17, 31).
- Li, Yan, Kourosh Khoshelham, Majid Sarvi, and Milad Haghani (2019a). “Direct generation of level of service maps from images using convolutional and long short-term memory networks.” In: *Journal of Intelligent Transportation Systems*, pp. 1–9 (cit. on pp. 16, 31, 46, 49).
- Li, Yan, Majid Sarvi, Kourosh Khoshelham, Milad Haghani, and Yuan Tian (2019b). *Multi-view crowd congestion map generation based on ensemble learning*. Tech. rep. (cit. on p. 43).
- Liciotti, Daniele, Marina Paolanti, Emanuele Frontoni, Adriano Mancini, and Primo Zingaretti (2016). “Person re-identification dataset with rgb-d camera in a top-view configuration.” In: *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*. Springer, pp. 1–11 (cit. on pp. 22, 26).
- Liu, An-An, Wei-Zhi Nie, Yu-Ting Su, Li Ma, Tong Hao, and Zhao-Xuan Yang (2015). “Coupled hidden conditional random fields for RGB-D human action recognition.” In: *Signal Processing* 112, pp. 74–82 (cit. on p. 27).
- Liu, Chunhui, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu (2017). “PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding.” In: *arXiv preprint arXiv:1703.07475* (cit. on pp. 23, 27).
- Liu, Jingchen, Robert T Collins, and Yanxi Liu (2011). “Surveillance camera autocalibration based on pedestrian height distributions.” In: *British Machine Vision Conference (BMVC)*. Vol. 2 (cit. on p. 15).
- Lo, BPL and SA Velastin (2001). “Automatic congestion detection system for underground platforms.” In: *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech*

- Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)*. IEEE, pp. 158–161 (cit. on pp. [12–14](#), [33](#), [45](#)).
- Loy, Chen Change, Ke Chen, Shaogang Gong, and Tao Xiang (2013). “Crowd counting and profiling: Methodology and evaluation.” In: *Modeling, Simulation and Visual Analysis of Crowds*. Springer, pp. 347–382 (cit. on p. [13](#)).
- Lv, Yisheng, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang (2015). “Traffic flow prediction with big data: a deep learning approach.” In: *IEEE Transactions on Intelligent Transportation Systems* 16.2, pp. 865–873 (cit. on p. [17](#)).
- Ma, Huadong, Chengbin Zeng, and Charles X Ling (2012). “A reliable people counting system via multiple cameras.” In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.2, p. 31 (cit. on p. [15](#)).
- Maddalena, Lucia, Alfredo Petrosino, and Francesco Russo (2014). “People counting by learning their appearance in a multi-view camera environment.” In: *Pattern Recognition Letters* 36, pp. 125–134 (cit. on p. [15](#)).
- Mathieu, Michael, Camille Couprie, and Yann LeCun (2015). “Deep multi-scale video prediction beyond mean square error.” In: *arXiv preprint arXiv:1511.05440* (cit. on p. [72](#)).
- Mousse, Mikaël A, Cina Motamed, and Eugène C Ezin (2017). “People counting via multiple views using a fast information fusion approach.” In: *Multimedia Tools and Applications* 76.5, pp. 6801–6819 (cit. on p. [15](#)).
- Munaro, Matteo, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti (2014a). “3D reconstruction of freely moving persons for re-identification with a depth sensor.” In: *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, pp. 4512–4519 (cit. on pp. [22](#), [26](#)).
- Munaro, Matteo, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool (2014b). “One-shot person re-identification with a consumer depth camera.” In: *Person Re-Identification*. Springer, pp. 161–181 (cit. on pp. [22](#), [26](#)).
- National Research Council (U.S.) (2010). *Highway capacity manual*. Washington, D.C: Transportation Research Board (cit. on pp. [11](#), [34](#)).
- Ni, Bingbing, Gang Wang, and Pierre Moulin (2011). “Rgb-d-huDaacT: A color-depth video database for human daily activity recognition.” In: *Computer Vision Workshops (ICCV Workshops), 2011*

- IEEE International Conference on.* IEEE, pp. 1147–1153 (cit. on p. 27).
- Ofli, Ferda, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy (2013). “Berkeley mhad: A comprehensive multimodal human action database.” In: *Applications of Computer Vision (WACV), 2013 IEEE Workshop on.* IEEE, pp. 53–60 (cit. on p. 27).
- Opitz, David and Richard Maclin (1999). “Popular ensemble methods: An empirical study.” In: *Journal of artificial intelligence research* 11, pp. 169–198 (cit. on p. 15).
- Oreifej, Omar and Zicheng Liu (2013). “Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 716–723 (cit. on p. 27).
- Pala, Federico, Riccardo Satta, Giorgio Fumera, and Fabio Roli (2016). “Multimodal person reidentification using RGB-D cameras.” In: *IEEE Transactions on Circuits and Systems for Video Technology* 26.4, pp. 788–799 (cit. on pp. 22, 26).
- Polikar, Robi (2006). “Ensemble based systems in decision making.” In: *IEEE Circuits and systems magazine* 6.3, pp. 21–45 (cit. on p. 15).
- Polus, Abishai, Joseph L Schofer, and Ariela Ushpiz (1983). “Pedestrian flow and level of service.” In: *Journal of transportation engineering* 109.1, pp. 46–56 (cit. on p. 13).
- Rahmani, Hossein, Arif Mahmood, Du Q Huynh, and Ajmal Mian (2014). “HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition.” In: *European conference on computer vision*. Springer, pp. 742–757 (cit. on p. 27).
- Rahmani, Hossein, Arif Mahmood, Du Huynh, and Ajmal Mian (2016). “Histogram of oriented principal components for cross-view action recognition.” In: *IEEE transactions on pattern analysis and machine intelligence* 38.12, pp. 2430–2443 (cit. on p. 27).
- Ristani, Ergys, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi (2016). “Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking.” In: *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking* (cit. on p. 84).
- Rokach, Lior (2010). “Ensemble-based classifiers.” In: *Artificial Intelligence Review* 33.1-2, pp. 1–39 (cit. on p. 15).
- Ryan, David, Simon Denman, Sridha Sridharan, and Clinton Fookes (2015). “An evaluation of crowd counting methods, features

- and regression models." In: *Computer Vision and Image Understanding* 130, pp. 1–17 (cit. on pp. 13, 14).
- Saleh, Sami Abdulla Mohsen, Shahrel Azmin Suandi, and Haidi Ibrahim (2015). "Recent survey on crowd density estimation and counting for visual surveillance." In: *Engineering Applications of Artificial Intelligence* 41, pp. 103–114 (cit. on p. 13).
- Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview." In: *Neural networks* 61, pp. 85–117 (cit. on p. 14).
- Seyfried, Armin, Bernhard Steffen, and Thomas Lippert (2006). "Basics of modelling the pedestrian flow." In: *Physica A: Statistical Mechanics and its Applications* 368.1, pp. 232–238 (cit. on pp. 12, 13).
- Seyfried, Armin, Bernhard Steffen, Wolfram Klingsch, and Maik Boltes (2005). "The Fundamental Diagram of Pedestrian Movement Revisited." In: *Journal of Statistical Mechanics: Theory and Experiment* (cit. on p. 13).
- Shahroudy, Amir, Jun Liu, Tian-Tsong Ng, and Gang Wang (2016). "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis." In: pp. 1010–1019. eprint: 1604.02808 (cit. on pp. 23, 27).
- Shirazi, Mohammad Shokrolah and Brendan Tran Morris (2017). "Looking at Intersections: A Survey of Intersection Monitoring, Behavior and Safety Analysis of Recent Studies." In: *IEEE Transactions on Intelligent Transportation Systems* 18.1, pp. 4–24 (cit. on pp. 23, 24).
- Si, Jianlou, Honggang Zhang, Chun-Guang Li, and Jason Kuen (2018). "Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification." In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 20).
- Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (cit. on pp. 14, 34, 47).
- Snoek, Cees GM, Marcel Worring, and Arnold WM Smeulders (2005). "Early versus late fusion in semantic video analysis." In: *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, pp. 399–402 (cit. on p. 16).
- Spinello, Luciano and Kai O Arras (2011). "People detection in RGB-D data." In: *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, pp. 3838–3843 (cit. on pp. 22, 26).

- Suh, Yumin, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee (2018). "Part-Aligned Bilinear Representations for Person Re-identification." In: *arXiv preprint arXiv:1804.07094* (cit. on p. 20).
- Sun, Yifan, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang (2018). "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)." In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 480–496 (cit. on p. 20).
- Sung, Jaeyong, Colin Ponce, Bart Selman, and Ashutosh Saxena (2011). "Human Activity Detection from RGBD Images." In: *plan, activity, and intent recognition* 64 (cit. on p. 27).
- Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks." In: *Advances in neural information processing systems*, pp. 3104–3112 (cit. on p. 14).
- Tan, Ben, Junping Zhang, and Liang Wang (2011). "Semi-supervised elastic net for pedestrian counting." In: *Pattern Recognition* 44.10, pp. 2297–2304 (cit. on p. 35).
- Tang, Nick C, Yen-Yu Lin, Ming-Fang Weng, and Hong-Yuan Mark Liao (2015). "Cross-camera knowledge transfer for multiview people counting." In: *IEEE Transactions on image processing* 24.1, pp. 80–93 (cit. on p. 15).
- Taylor, Michael AP, Jeremy E Woolley, and Rocco Zito (2000). "Integration of the global positioning system and geographical information systems for traffic congestion studies." In: *Transportation Research Part C: Emerging Technologies* 8.1-6, pp. 257–285 (cit. on p. 70).
- Tripathi, Gaurav, Kuldeep Singh, and Dinesh Kumar Vishwakarma (2019). "Convolutional neural networks for crowd behaviour analysis: a survey." In: *The Visual Computer* 35.5, pp. 753–776 (cit. on p. 4).
- Villegas, Ruben, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee (2017). "Decomposing motion and content for natural video sequence prediction." In: *ICLR* 1.2, p. 7 (cit. on p. 18).
- Virgona, Alexander, Nathan Kirchner, and Alen Alempijevic (2015). "Sensing and perception technology to enable real time monitoring of passenger movement behaviours through congested rail stations." In: 1.October, pp. 1–14 (cit. on p. 21).
- Wang, Jiang, Zicheng Liu, Ying Wu, and Junsong Yuan (2012). "Mining actionlet ensemble for action recognition with depth

- cameras." In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, pp. 1290–1297 (cit. on p. 27).
- Wang, Jiang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu (2014a). "Cross-view action modeling, learning and recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649–2656 (cit. on pp. 23, 27).
- Wang, Jinqiao, Wei Fu, Jingjing Liu, and Hanqing Lu (2014b). "Spatiotemporal group context for pedestrian counting." In: *IEEE Transactions on Circuits and Systems for Video Technology* 24.9, pp. 1620–1630 (cit. on p. 35).
- Wang, Keze, Xiaolong Wang, Liang Lin, Meng Wang, and Wang-meng Zuo (2014c). "3d human activity recognition with reconfigurable convolutional neural networks." In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pp. 97–106 (cit. on pp. 23, 27).
- Wang, Xiaogang (2013). "Intelligent multi-camera video surveillance: A review." In: *Pattern recognition letters* 34.1, pp. 3–19 (cit. on p. 59).
- Wei, Ping, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu (2013). "Modeling 4d human-object interactions for event and object recognition." In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3272–3279 (cit. on p. 27).
- Wolpert, David H (1992). "Stacked generalization." In: *Neural networks* 5.2, pp. 241–259 (cit. on p. 16).
- Wu, Yuankai and Huachun Tan (2016). "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework." In: *arXiv preprint arXiv:1612.01022* (cit. on p. 17).
- Xia, Lu, Chia-Chih Chen, and J. K. Aggarwal (2012). "View Invariant Human Action Recognition Using Histograms of 3D Joints." In: *Workshop on Human Activity Understanding from 3D Data*, pp. 20–27 (cit. on p. 27).
- Xingjian, SHI, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo (2015). "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." In: *Advances in neural information processing systems*, pp. 802–810 (cit. on pp. 17, 63, 64).
- Yang, Liu and Rong Jin (2006). "Distance metric learning: A comprehensive survey." In: *Michigan State University* 2.2, p. 4 (cit. on p. 19).

- Yogameena, B and C Nagananthini (2017). "Computer vision based crowd disaster avoidance system: A survey." In: *International journal of disaster risk reduction* 22, pp. 95–129 (cit. on p. 4).
- Yu, Bing, Haoteng Yin, and Zhanxing Zhu (2017). "Spatio-temporal Graph Convolutional Neural Network: A Deep Learning Framework for Traffic Forecasting." In: *arXiv preprint arXiv:1709.04875* (cit. on p. 17).
- Yu, Gang, Zicheng Liu, and Junsong Yuan (2014). "Discriminative orderlet mining for real-time recognition of human-object interaction." In: *Asian Conference on Computer Vision*. Springer, pp. 50–65 (cit. on p. 27).
- Yu, Haiyang, Zhihai Wu, Shuqin Wang, Yunpeng Wang, and Xiaolei Ma (2017a). "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks." In: *Sensors* 17.7, p. 1501 (cit. on p. 17).
- Yu, Qian, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales (2017b). "The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching." In: *arXiv preprint arXiv:1711.08106* (cit. on pp. 20, 82).
- Zaki, M.H. and T. Sayed (2017). "Automated Analysis of Pedestrian Group Behavior in Urban Settings." In: *IEEE Transactions on Intelligent Transportation Systems* 19.6, pp. 1880–1889 (cit. on p. 18).
- Zhang, Cha and Yunqian Ma (2012). *Ensemble machine learning: methods and applications*. Springer (cit. on p. 47).
- Zhang, Junbo, Yu Zheng, and Dekang Qi (2017). "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction." In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pp. 1655–1661 (cit. on pp. 16, 17).
- Zhang, Zhimeng, Jianan Wu, Xuan Zhang, and Chi Zhang (2017). "Multi-target, multi-camera tracking by hierarchical clustering: recent progress on DukeMTMC Project." In: *arXiv preprint arXiv:1712.09531* (cit. on p. 82).
- Zheng, Liang, Yi Yang, and Alexander G Hauptmann (2016). "Person re-identification: Past, present and future." In: *arXiv preprint arXiv:1610.02984* (cit. on p. 19).
- Zheng, Zhedong, Liang Zheng, and Yi Yang (2017). "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro." In: *Proceedings of the IEEE International Conference on Computer Vision* (cit. on p. 84).

Zitouni, M Sami, Harish Bhaskar, J Dias, and Mohammed E Al-Mualla (2016). "Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques." In: *Neurocomputing* 186, pp. 139–159 (cit. on p. [4](#)).