

Department of Computing

**A Spatio-temporal Learning Approach for Crowd Activity  
Modelling to Detect Anomalies**

Arjun Rao

This thesis is presented for the Degree of  
Master of Science (Computer Science)  
of  
Curtin University of Technology

November 2009

## **Declaration**

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: .....

Date:.....

Dedicated to my beloved Mother and Father...

## Abstract

With security and surveillance gaining paramount importance in recent years, it has become important to reliably automate some surveillance tasks for monitoring crowded areas. The need to automate this process also supports human operators who are overwhelmed with a large number of security screens to monitor. Crowd events like excess usage throughout the day, sudden peaks in crowd volume, chaotic motion (obvious to spot) all emerge over time which requires constant monitoring in order to be informed of the event build up. To ease this task, the computer vision community has been addressing some surveillance tasks using image processing and machine learning techniques. Currently tasks such as crowd density estimation or people counting, crowd detection and abnormal crowd event detection are being addressed. Most of the work has focused on crowd detection and estimation with the focus slowly shifting on crowd event learning for abnormality detection.

This thesis addresses crowd abnormality detection. However, by way of the modelling approach used, implicitly, the tasks of crowd detection and estimation are also handled. The existing approaches in the literature have a number of drawbacks that keep them from being scalable for any public scene. Most pieces of work use simple scene settings where motion occurs wholly in the near-field or far-field of the camera view. Thus, with assumptions on the expected location of person motion, small blobs are arbitrarily filtered out as noise when they may be legitimate motion in the far-field. Such an approach makes it difficult to deal with complex scenes where entry/exit points occur in the centre of the scene or multiple pathways running from the near to the far-field of the camera view that produce blobs of differing sizes. Further, most authors assume the number of directions people motion should exhibit rather than discover what these may be. Approaches with such assumptions would result in loss of accuracy while dealing with (say) a railway platform which shows a number of motion directions, namely two-way, one-way, dispersive, etc. Finally, very few contributions of work use time as a video feature to model the human intuitiveness of time-of-day abnormalities. That is certain motion patterns may be abnormal if they have not been seen for a given time of day. Most works use it (time) as an extra qualifier to spatial data for trajectory definition.

In this thesis most of these drawbacks have been addressed by dealing with these in the modelling of crowd activity. Firstly, no assumptions are made on scene structure or blob sizes resulting therefrom. The optical flow algorithm used is robust and even the noise presented (which is in fact unwanted motion of swaying hands and legs as opposed to that from the torso) is fairly consistent and therefore can be factored into the modelling. Blobs, no matter what the size are not discarded as they may be legitimate emerging motion in the far-field. The modelling also deals with paths extending from the far to the near-field of the camera view and segments these such that each segment contains self-comparable fields of motion. The need for a normalisation factor for comparisons across near and far field motion fields implies prior knowledge of the scene. As the system is intended for generic public locations having varying scene structures, normalisation is not an option in the processing used and yet the near & far-field motion changes are accounted for. Secondly, this thesis describes a system that learns the true distribution of motion along the detected paths and maintains these. The approach is such that doing so does not generalise the direction distributions which would cause loss in precision. No impositions are made on expected motion and if the underlying motion is well defined (one-way or two-way), then this is represented as a well defined distribution and as a mixture of directions if the underlying motion presents itself as so. Finally, time as a video feature is used to allow for activity to re-enforce itself on a daily basis such that motion patterns for a given time and space begin to define themselves through re-enforcement which acts as the model used for abnormality detection in time and space (spatio-temporal). The system has been tested with real-world data datasets with varying fields of camera view. The testing has shown no false negatives, very few false positives and detects crowd abnormalities quite well with respect to the ground truths of the datasets used.

## Acknowledgements

This thesis is the culmination of input from a number of people in the form of financial support, strength, mentors, beers on steps, escapist policies, long walks, etc. I owe thanks and feel grateful firstly to my parents, my brother and sister for their support and encouragement.

At the department of Computing my good mates, Thorsten, Min, Annika, Thomas, Simon and Daniel have always been there for brainstorming and exchanging ideas. They and others participated in data collection exercises that were invaluable. I'd like to thank Patrick, our senior post doc. and now lecturer for imparting the thought structures for appreciation of good/reasonable methodology in research. I believe I have made the decisions I have while modelling this thesis based on the thinking culture developed with my interaction with Patrick and my supervisors.

My supervisors, Dr. Tele Tan and Dr. Mihai Lazarescu not only helped me pick the topic for my thesis, but always passed on that little hint which would get me through my moments of methodology uncertainty. Mihai would talk at length about the perfection we all strive for and how one must strike a balance between accepting a practical result vs chasing a perfect one. Tele always had an uncanny way of coming to my rescue when I saw nothing but a wall in front of me. And most of all, they have been most patient with me and I am sincerely grateful for that.

Finally, I would like to thank my Chairperson Prof. Geoff West and Dr. Svetha Venkatesh who along with Mihai encouraged me to undertake a masters' degree by research from which I feel now that I have the mental tools and tenacity to solve any problem that might come my way.

A word of thanks to the tech staff at Computing, my housemates who are great chefs and my friends here in Perth who invite me for dinner where I once again experience tastes from back home.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Approach . . . . .	3
1.2	Significance . . . . .	4
1.3	Aims . . . . .	6
1.4	Thesis Outline . . . . .	6
<b>2</b>	<b>Background</b>	<b>8</b>
2.1	Introduction . . . . .	8
2.2	Brief introduction to Image Processing & Computer Vision techniques . . . . .	10
2.2.1	Features & Feature Extraction techniques . . . . .	12
2.3	The tracking problem . . . . .	16
2.4	Related work in crowd analysis . . . . .	19
2.5	Summary . . . . .	24
<b>3</b>	<b>Methodology</b>	<b>25</b>
3.1	Introduction . . . . .	25
3.2	Setting the scene . . . . .	26
3.3	Preliminaries . . . . .	26
3.4	System Overview . . . . .	32
3.5	Feature Extraction . . . . .	34
3.6	3D Spatio-temporal Label Generation . . . . .	35
3.6.1	(i) Generating/Learning Path Labels . . . . .	36
3.6.2	(ii) Generating Temporal Labels: . . . . .	38
3.7	Model Generation for activity representation . . . . .	38

3.8	Model Comparison . . . . .	43
3.9	Summary . . . . .	45
<b>4</b>	<b>Datasets &amp; Experiments</b>	<b>46</b>
4.1	Introduction . . . . .	46
4.2	Datasets . . . . .	47
4.2.1	Dataset 1: Curtin university pathways . . . . .	48
4.2.2	Dataset 2: Perth train station passageway . . . . .	51
4.3	Experiments and Ground-truthing . . . . .	56
4.3.1	Experiment 1: Testing path learning . . . . .	56
4.3.2	Results of path detection: University sample . . . . .	62
4.3.3	Results of path detection: Railway Passageway dataset . . . . .	64
4.3.4	Experiment 2: Validation of proposed activity modelling and suitability of temporal resolution . . . . .	66
4.3.5	Determination of Temporal Resolution . . . . .	67
4.3.6	Parameters and settings used . . . . .	73
4.3.7	Results: Comparison of similar samples . . . . .	75
4.3.8	Results: Comparison of dissimilar samples . . . . .	76
4.3.9	Experiment 3: System performance by way of test sample classification with the trained model . . . . .	78
4.3.10	Parameters and settings used . . . . .	79
4.4	A Discussion of Weaknesses of the System . . . . .	83
4.4.1	Single person anomalies . . . . .	83
4.4.2	A group of people standing still . . . . .	83
4.4.3	Anomalies across discrete time windows . . . . .	84
4.4.4	Minimum time needed to detect an anomaly . . . . .	84
4.5	Summary . . . . .	84
<b>5</b>	<b>Conclusion</b>	<b>86</b>
5.1	Future Work . . . . .	87
	<b>Bibliography</b>	<b>93</b>



# List of Algorithms

- 1 Heuristics to classify outcome of summary comparison as normal or abnormal. 44

# List of Figures

3-1	Motion in a 3D setting . . . . .	27
3-2	System Overview . . . . .	33
3-3	Activity representation for a given time period . . . . .	39
3-4	Generated model of activity representation from training data . . . . .	41
4-1	Empty pathways at a University . . . . .	49
4-2	Examples of Sparse and Dense usage of pathways at the University. . . . .	49
4-3	3D scatter plots of activity spread over 60 minutes (Z axis) across the scene. The different colours indicate different direction bins. Each path presents roughly two colours above it indicating two-way motion. . . . .	50
4-4	Sparse and dense usage of railway passageway. . . . .	51
4-5	3D scatter plots of activity spread over time across the scene. The plots have been oriented so as to show the length of the passageway. With reference to figure 4-4, the green and blue colour represents the bulk of the people moving from the far-field to the near-field making their way to the outside of the station while the mixture of red and yellow represents people moving from the near-field to the far-field, that is towards the underground train platform. . . . .	53
4-6	Images of abnormal event staging . . . . .	55
4-7	Ground truthed paths for the two datasets. . . . .	58
4-8	Results for experiment 1 (Path detection): University Pathways . . . . .	63
4-9	Results for experiment 1 (Path Detection): Railway Passageway. As can be seen re-enforcing motion from (b), (c) and (d) provides more detail in (e). . . . .	65
4-10	Ground truthing example of a similar pair of samples . . . . .	68
4-11	Ground truthing example of a dissimilar pair of samples . . . . .	69

4-12 Comparison decomposition of similar samples - 1 . . . . .	74
4-13 Comparison decomposition of similar samples - 2 . . . . .	75
4-14 Comparison decomposition of dissimilar samples - 1 . . . . .	76
4-15 Comparison decomposition of dissimilar samples - 2. . . . .	77

# List of Tables

2.1	Tracking Categories adapted from a survey paper on state of the art algorithms (Yilmaz, Javed, and Shah 2006) . . . . .	18
3.1	Example of a text file containing motion information . . . . .	35
4.1	List of videos samples used of the train station passage . . . . .	52
4.2	Path Detection: Empirical summary of binning scheme for direction descriptor	59
4.3	Path Detection: Empirical summary of binning scheme for speed descriptor .	60
4.4	Samples used for path learning in Experiment 1 . . . . .	61
4.5	Summary comparisons using a resolution of 3 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar . . . . .	71
4.6	Summary comparisons using a resolution of 10 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar . . . . .	72
4.7	Summary comparisons using a resolution of 15 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar . . . . .	72
4.8	Summary comparisons using a resolution of 20 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar . . . . .	72
4.9	Samples used for experiment 2 . . . . .	72
4.10	Motion histogram settings for experiments 2 and 3 . . . . .	73
4.11	Normalised cross-correlated values of the sample comparisons, $\delta_\theta$ , $\delta_\nu$ and $\delta_\rho$ .	73
4.12	Samples used in experiment 3 (Railway passage only) . . . . .	78
4.13	Results of normal test samples with the trained model . . . . .	81
4.14	Results of abnormal test samples with the trained model . . . . .	82

# Chapter 1

## Introduction

Today video surveillance is commonly found in most forms of security setups especially those for public locations such as bus and train stations. For such setups it is necessary to store recorded footage for a number of days for reasons such as investigating crime, monitoring for unusual behaviour as well as general activity. The conventional way to inspect video for different events is to manually fast forward through tens of hundreds of hours of video in search of some crowd abnormality, for example crime or commotion. The manual search process could (a) take large amounts of time to find abnormalities, (b) be error prone due to fatigue faced by the human operator, (c) time sensitive results could be delayed by weeks. The need to reliably automate the search process for retrieving crowd anomalies such as emergencies and irregular crowd volumes from stored video is evident as doing so would assist surveillance teams by **primarily saving time**. Further, algorithmic processing has the ability to process large samples of video and provide **summaries** which would enable security teams to take informed decisions that may have life saving outcomes.

Searching or querying video for events is synonymous to queries to textual databases such as web search engines. In the case of textual databases, the unit of comparison is discrete and well defined, namely, the word. A number of words make a sentence or a phrase with higher semantic meanings which would return accurate results from the database. Unfortunately, the unit of comparison in video and images is somewhat less well defined or unstructured, namely image features such as colour, texture, etc. The search process in a video database

pivots around tracking people and the information derived supports higher semantics such as path detection. Hence all queries depend on successfully tracking regions of colour (implying people) or some feature across image frames. Deriving scene semantics as queries to a video data-store in computer vision is a hierarchical process, that is, intermediate results build upon initial results and further results build upon intermediate results. Tracking colour or any other image feature in video is synonymous to searching for occurrences of text in a paragraph or some body of text. However, such a seemingly simple task of colour or feature matching across video frames as opposed to text matching across paragraphs is rendered relatively less successful due to (a) industry standard video implying low video resolution, low frame rate at typically 3 to 10 fps, (b) noise and colour jitter presented by the environmental changes such as cloud cover and swaying trees and (c) minor electrical disturbances which are some of the challenges faced during feature matching. Another problem in Vision is the problem of object occlusion that clearly hinders image feature tracking. Heavily crowded locations such as train and bus stations present severe occlusion which when coupled with poor video quality significantly reduces the chances of deriving successful trajectory tracks (matching features across images) of image features for higher level analysis.

The general goals of most security setups and of this thesis are to monitor crowd activity at public locations for anomalies such as crowd bottlenecks, commotion, untimely or unexpected volumes of people, deriving usage statistics and being informed of the the way a location is used throughout the day. Queries answering such requirements may be broadly classified into two categories, (1) person oriented queries and (2) crowd or groups of people oriented queries. Using the tracking approach agglomeratively, one could perform analysis on individuals, then groups of people and then crowds. However, this would require an overhead vantage point as otherwise the problem of occlusion makes it difficult to avoid broken trajectories. Furthermore, severe occlusion results in a large number of broken trajectories which effectively resembles tracking features between two frames. In this thesis, the worst case scenario of occlusion is considered and the observation that extremely broken trajectories are simply a collection of motion vectors representing the motion fields of people movements. Therefore in this thesis, motion fields are examined to learn patterns and not trajectory analysis. This thesis adopts the second category from above, that is, crowd movements analysis by way of

motion field analysis. Hence, fine grain events or person events lie outside the scope of this thesis as it concerns itself with crowd movement analysis.

Abnormalities specific to public locations gain their specific spatial boundary by the space where they occur (in the video scene) and their temporal boundary from the time of day at which they occur. It is desirable to acquire this information as it would not only time index the video but also pin point the spatial location in the scene where the abnormality occurred. The task of modelling activity therefore is spatio-temporal in nature (considers both space and time) which makes such a system practical for it's purpose of surveillance.

## 1.1 Approach

The worst case scenario of occlusion is difficult to deal with using trajectory analysis and hence that methodology of analysis is avoided. In this thesis the features used are foreground motion and time which are simply motion fields created by image feature propagation from one image frame to the next for a given sample size. Therefore the entire video may be thought of as a video of motion fields with no real knowledge of single person descriptions. Motion fields perceptively distinguish themselves from one another based on their properties which have been observed as a mixture of directions having a mixture of speeds and an observed density. Simple observation of motion fields reveal these three properties as the distinguishing characteristics of a field. It is the properties of motion that enable the approach to learn spatial variations in the scene which helps in path detection and learning localised definitions of activity on a per path basis. The idea of learning properties of a motion field is the facilitating mechanism to achieve spatial learning while the temporal separation of activity provides the ability of time-of-day specific activity learning.

The fundamental idea for generating spatial and temporal definitions is that of *re-enforcement* inspired from the idea of density maps used for obtaining amount of scene usage over time from (Velastin et al. 2004). Given a 3D motion mass representing the collection of movements across an image scene and over a given time span, the **vertical re-enforcements** or vertically

collapsing the mass of motion reveals spatial boundaries while **horizontal re-enforcements** across a number of days, that is, superimposing a number of 3D motion masses each of which were obtained from a single day reveals activity trends and variations over the sample size (temporally). Activity summaries bounded by spatial and temporal bounds as explained in Chapter 3 are maintained from these horizontal and vertical re-enforcements which is representative of the desired model body of spatio-temporal activity summaries for comparisons. The summarisation process reduces the millions of motion vectors into compact representations via the properties of motion mentioned above. Such an approach not only simplifies processing of the model generation but also the comparison process. The method is effective as demonstrated in the results in Chapter 4. Most importantly, while generating the model, motion fields are not restricted to being defined by a single representative direction, but are represented as a mixture of directions. Hence if the motion field is well defined, the mixture of directions would show a well defined structure and if the motion field were relatively less well defined, then the same would be reflected in the mixture representing it. That is, no assumptions are made on motion fields during the summarisation process.

## 1.2 Significance

The significance of the proposed approach in Chapter 3 is as follows:

- The main significance of this thesis is the ability to deal with high occlusion locations (worst case scenarios included). This is because tracking is avoided and the problem is redefined as one of learning the properties of motion fields.
- Abnormalities are classified so based on not just spatial irregularities but also if they occurred at an unexpected time-of-day. Such classifications are possible by considering both space and time during the training phase.
- The proposed approach considers that changes in motion may be due to (a) scene structure and (b) perspective changes, that is the near-field/far-field effect in the camera view. Therefore the modelling process incorporates localised spatial definitions based



on motion similar regions and *these regions need not be the same as the observed paths* (refer section 3.3). This approach is adopted because the system is intended to be generic, implying no prior scene information is available and hence no normalising factor may be derived for noise removal or intra-scene comparisons. Moreover, it would be inappropriate to apply a single global threshold for comparisons or noise removal in real world scenes that present differing depths of scene. Such considerations render the proposed approach scalable for most real world scene settings.

The system described in this thesis does not make assumptions on:

- Blob sizes or far-field motion as noise. A number of polar plots of motion showed that noise in the sequences were consistent in quantity and relatively less compared to legitimate motion, therefore motion vectors representing noisy directions were not thresholded. However, those with large displacement values showing large deviations from the mean were discarded.
- Again, no assumptions are made on the direction of people flow and no attempt is made to search for a well defined or representative motion direction while processing (refer Chapter 3). Instead, the whole distribution is used as being representative of either one-way motion or two-way motion or a mixture of directions representing the true distribution of motion being witnessed along a given path.
- After using motion analysis of true distributions for path detection, this thesis uses time to build definition of activity throughout the day (or specified time period) on a path-wise basis. Such an approach provides realistic time-of-day abnormality detection, path detection, implicit knowledge of motion estimation over a period of time and not just a snapshot of data.

## 1.3 Aims

The aims needed to fulfil the goal of this thesis which is detecting abnormalities within crowd activity using motion analysis are as follows:

- To extract foreground motion and time as features from surveillance video.
- To generate a model representation of activity for a given time span such as five hours or perhaps a day with sufficient variations such that the model is truly representative of daily or weekly activity for that span of time. Moreover, in order to render the system as scalable, the modelling process should take into account spatial changes in motion fields caused due to the way people use a scene or the near and far-field effect in the camera view that introduces perspective changes in motion displacements and directions. This sub-aim may be summarised as spatio-temporal modelling of activity.
- To devise a method to compare test video samples against the trained model in order to classify the test sample as normal or abnormal in a way such that it is clear where in the scene (spatially) and within which span of time within the test sample (temporally) did an abnormality occur, if any.

## 1.4 Thesis Outline

Chapter 1 (this) explained the requirement, that is crowd flow analysis and introduced the general approach and aims of the thesis. Chapter 2 provides summaries of related and supporting work in the field of computer vision. Topics such as background subtraction and optic flow that are used in image processing and feature extraction are discussed. Also summaries of a number of recent techniques in crowd analysis are presented.

Chapter 3, the methodology, introduces the terms, notations used, modelling approach and the system implementation. The system implementation entails video feature extraction,

processing, training the model and the method for comparing a test sample of activity with the trained model.

Chapter 4 explains the datasets used, experiments and their corresponding ground truthing methods and results. Finally Chapter 5 concludes this thesis emphasising the strengths of the approach used validated by the results and suggestions as to how to improve the proposed approach for activity modelling.

## Chapter 2

# Background

### 2.1 Introduction

Problems occur in crowds each year in public locations such as train stations, stadiums and bus terminals. At times the nature of the crowd anomaly may not be grasped by a human operator as abnormalities take time to build up for which the surveillance operator needs to constantly observe the footage. Abnormalities may be of different types, for example, excess usage over a number of hours may go undetected by a surveillance operator as he is seeing just a snapshot of the activity. On the other hand abnormalities such as commotions, chaotic movements and bottlenecks may be more obvious to spot. Further, if a single camera's view covers a large area such as a number of railway platforms, it is reasonable for surveillance operators to spot general abnormalities along each platform while missing out on the slow emerging ones. This thesis addresses, the need to automate the process of crowd abnormality detection by taking into account spatial locality and temporal build-up of activity so as to gain a reasonable measure of crowd activity before comparing it against a trained model which itself was generated using time-space information. Foreground motion and time are the only working features and the approach devised does not make any location/scene specific assumptions as it is intended to be generically used for any public location.

In computer vision, the majority of the work in crowd analysis has focused on people count-

ing or density estimation as opposed to event mining for abnormalities. This view is also observed in a survey paper by [Zhan et al. \(2008\)](#). The early work of [Boghossian and Velastin \(1999\)](#) used optical flow to detect motion and searched for specific patterns relating to specific abnormalities (further described in section 2.4). Recent work by [Andrade et al. \(2006\)](#) trains multi-object Hidden Markov Models (HMMs) for detecting a specific crowd emergency event such as bottlenecks. Most approaches have typically used: (a) snapshots of data, that is low temporal resolution to classify crowd types rather than higher temporal resolution, that is, a larger window of time that is more likely to bring a crowd event into focus, (b) work with simplistic scenes containing one or two paths. With simple scenes containing similar object sizes and similar activity/motion trends across the scene, the authors use global thresholds applicable for the whole scene. Such an approach renders previous work unfit for video scenes containing a number of pathways in the near and far-field of the camera view causing differing object/blob sizes that would require more than one global threshold for comparisons. For realistic scene settings such as railway platforms and bus terminals containing differing camera fields, accounting for spatial differences in motion patterns seems desirable as differing sizes of people caused by near and far-field effect would use pathways in differing degrees of density based on the time of day. Therefore, a threshold for amount of activity (say) for one path may be highly lacking for another. Further, one railway platform may experience quick and dispersive motion during peak hours while another may remain relatively unused. Therefore learning local spatial patterns is desirable as it can help significantly in anomaly detection.

In addition, it is important to qualify regions in a scene (space) with time associated activity levels. Realistically one would not expect to see the same levels of activity in a scene throughout the day. For example, during peak hours, activity runs high while during off-peak hours it runs low or moderate. Therefore it is desirable to have a model of activity that not only accounts for spatial changes in activity patterns but also lists different activity types based on the time-of-day. Most of the work in the existing literature use time as a third dimension for collecting trajectory information only ([Brostow and Cipolla 2006](#); [Rabaud and Belongie 2006](#); [Wang, Tieu, and Grimson 2006](#)). Incorporating a temporal qualification to the levels of activity occurring adds a new level to classifying activity as abnormal based on

the time-of-day. For example, it is abnormal to observe peak-hour levels of activity during lunch-time as they are generally seen in the mornings and evenings.

The significance of the problem is noted as incorporation of time qualification to activity levels and accounting for spatial changes in motion patterns that are naturally effected by camera perspective and complex scenes. Modelling activity in such a fashion is referred to as *modelling activity in a time-space fashion*. These incorporations make the approach scalable for any public location with complex scene semantics and abnormalities are detected based on time of day and not just snapshots of data (low temporal resolution). The rest of this chapter proceeds thus – a brief introduction to image processing and computer vision followed by the tracking problem and how it limits the scope of processing that makes it unfit for crowd processing. Finally, related work in crowd analysis will be discussed followed by a brief summary.

## **2.2 Brief introduction to Image Processing & Computer Vision techniques**

This brief introduction provides an overview of computer vision, the steps involved in a generic pattern recognition system, a note on features and some feature extraction techniques.

Computer vision deals with techniques for extracting image frames with additional processing from a visual sensor such as a still camera or video camera. The further processing involves image processing techniques among others such as image restoration, smoothing, image feature extraction such as colour, texture, edge information, etc. In industrial applications, computer vision is used for visual quality control checks for items such as automobile brake assemblies (<http://www.avalonvisionsolutions.com>) and in some consumer goods such as digital cameras where product features such as face detection and enhancements come standard. Work still under progress includes surveillance tasks such as face recognition, person tracking, action recognition and event detection in crowds.

Machine learning is the study of learning patterns of data from a visual or other sensor or from a database. The database may also be a collection of image features extracted from the video sensor using computer vision feature extraction techniques. Machine learning and pattern recognition are closely linked as pattern recognition mainly deals with algorithms that assign class labels to candidate features (potential patterns) while the science behind designating the class label is dealt with by machine learning, namely supervised and unsupervised learning amongst others.

Given the foregoing, today's computer vision systems are a combination of computer vision, machine learning and pattern recognition techniques used to solve problems such as (a) surveillance tasks such as tracking or counting people, finding paths in a scene, (b) recognition and synthesis of vehicle number plates, (c) medical imaging, (d) analysis of high dimensional data extracted from a number of sensors and many more real world problems. As all these systems including the one described in this thesis are essentially signal processing systems, they generally tend to have similar processing steps. A brief outline of the processing steps for a generic pattern recognition system is provided next.

- i Pre-processing:** This phase removes noise from the video signal (or simply video) introduced by factors such as electrical disturbance, light changes due to environmental changes, etc. This is commonly carried out by image smoothing or filtering with a median, mean or Gaussian filter. The overall objective is to obtain a cleaner, smoother signal for feature extraction.
- ii Feature extraction:** The phase of extracting features from the signal. In the case of an image or video signal, the features in question would be colour, texture, time (frame number) or foreground motion (complex feature).
- iii Processing:** The extracted features are either reduced in dimensionality using techniques like Principal Component Analysis (PCA) for further efficient analysis or grouped/clustered based on similarity via a distance measure such as Euclidean so as to learn different class labels present within the data. The process of class label discovery may be supervised or unsupervised.

**iv Higher level processing:** Some systems may not contain this step or may use this phase for classification of test samples with the learned model. Other systems may build upon the output of the previous step, for example, if the previous step involved detecting paths in a scene, then this step may involve learning event and event labels along the detected paths.

### 2.2.1 Features & Feature Extraction techniques

In this subsection, features with respect to image & video processing and some feature extraction techniques will be discussed.

Features are image elements such as colour, texture and edge/line information. Features may be grouped based on some similarity measure in order to build object definitions within an image. The process of identifying image objects or moving ones in the case of video is known as segmentation. In video sequences, the object of interest which is moving is referred to as a *blob* while the entirety of moving objects is referred to as the foreground and by contrast all that does not move is referred to as the background. Image or video segmenting is however performed after the process of feature extraction via a number of techniques such as (a) edge detection algorithms such as Sobel (Sobel and Fieldman 1968) and Canny (Canny 1986). (b) Local binary patterns (Heikkila et al. 2004) (which is texture driven), (c) segmenting or detecting the presence of blobs in video sequences is performed by a process called background subtraction (explained later), (d) region labelling or region growing and additional processes in the literature if required.

Image and video features may be broadly classified as simple and complex features respectively. Simple features as mentioned above are colour, texture, points and edges while complex features are those that may be decomposed into simple ones; they are motion, blobs, etc. Motion may be defined as displacement of a single point or patch (group of points) from one frame to another while a blob may be distinguished as a continuous region (primarily defined by continuous colour or texture). A blob in video sequences carries the connotation that



the same blob must be tracked across video frames for surveillance reasons known as **feature correspondence** or simply tracking. Tracking an image feature or a region or a blob across frames involves identifying candidate points or regions in two consecutive image frames and then determining them to be similar by some comparison process such as correlation or differencing their values against a user-defined threshold of some distance measure such as Euclidean. Comparisons for colour regions may be done using colour histograms (Zivkovic and Krose 2004) for which there are a number of distance measures such as the Bhattachariya distance, the earth mover distance (Rubner et al. 2000) or cross correlation.

The video features used in this thesis are foreground motion and time. The foreground motion is commonly obtained using a combination of background subtraction and optical flow. Background subtraction provides the foreground image regions where motion of interest occurs and this is used as a filtering tool to select only the foreground optical flow while discarding the rest. Brief descriptions of background subtraction and optical flow are provided below.

### 2.2.1.1 Optic flow

In the case of video sequences optic flow is the motion/displacement of a point (pixel colour) or group of points from one image frame to another. A number of techniques perform this computation, one of which is (Lucas and Kanade 1981; Black and Anandan 1993) and most of these depend upon the pioneering work of Horn and Schunck (Horn and Schunck 1980). By computing motion fields the following is possible (a) motion estimation (Torr and Zisserman 1999; Irani and Anandan 1999) (b) object reconstruction from optic flow given fixed camera and lighting conditions with impressive results (Zhang et al. 2003), (c) object segmentation by motion, (d) recognising action at a distance by using optic flow template maps for image frame classification (Efros et al. 2003). Optic flow based on the technique by Bouguet (2000) was used in this thesis.

### 2.2.1.2 Background subtraction

The goal of background subtraction is to reveal or segment the foreground moving regions. These regions are usually representative of people or moving objects. The most basic principle is to apply frame differencing between two consecutive frames, the result of which reveals objects that have moved (the foreground) while reducing all other values to zero in the differencing operation - the background or parts of the image that have not moved. This simplistic approach lacks some desirable capabilities such as (1) being a simplistic differencing approach, a person standing still even for a second would disappear into the background, (2) the process would introduce false positives or noise in the form of swaying trees and changes in light due to passing clouds. One of the more popular and reliable background subtraction techniques that learned the foreground and modelled the above two concerns in the background was presented by Stauffer and Grimson in (Stauffer and Grimson 2000) and is the one used in this thesis. It is also regarded as having very good modelling accuracy by Piccardi (2004) and McIvor (2000) both of whom conduct a survey of various background subtraction techniques.

Stauffer and Grimson (2000) model the recent history of colour values for each pixel by maintaining  $K$  Gaussian distributions for each pixel. Each of the  $K$  Gaussians for a given pixel  $(x_0, y_0)$  is weighted based on persistence of colour values. The scheme of  $K$  Gaussians thus allows a distribution for a long time persisting background colour, one for temporarily still objects/people and a couple for transient colours or moving objects. Given a history of image values for pixel  $(x_0, y_0)$  from time 1 to  $t$ ,  $X_1, X_1, \dots, X_t$ , the probability of current value for the pixel is

$$P(X_t) = \sum_{i=1}^K \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \quad (2.1)$$

where,

$K$  is the number of distributions

$\omega_{i,t}$  is the estimate of the weight of the  $i^{th}$  Gaussian at time  $t$

$\mu_{i,t}$  is the mean of the  $i^{th}$  Gaussian at time  $t$

$\Sigma_{i,t}$  is the covariance matrix of the  $i^{th}$  Gaussian at time  $t$

$\eta$  is a Gaussian probability density function and

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu)^T \Sigma^{-1} (X_t - \mu)} \quad (2.2)$$

A pixel value is compared against  $K$  distributions until a match is found (2.5 standard deviations). If there is no successful match, then the least probable distribution is updated with its mean as the pixel value with large variance and an initial low weight. The update equation is given by

$$\omega_{k,t} = (1 - \alpha)\omega_{k,t} + \alpha(M_{k,t}) \quad (2.3)$$

where,

$M_{k,t}$  is 1 for the matched model and 0 for the unmatched ones  $\alpha$  is the learning rate

After each update operation, the weights are re-normalised. The background is chosen from the first  $B$  distributions given by  $B$  after ordering them based on  $\omega/\sigma$  and  $B$  is given by,

$$B = \operatorname{argmin}_b \left( \sum_{k=1}^b \omega_k > T \right) \quad (2.4)$$

The contribution of Stauffer and Grimson show us a number of things about system evolution, namely (a) localised models instead of a single global model. In this case it was pixel-based. (b) Learning - there was no hard threshold, but these (thresholds) were learnt by learning the foreground and the background. (c) and lastly the consideration of time, that is, maintaining historical evidence per pixel.

## 2.3 The tracking problem

In computer vision, surveillance tasks such as action recognition, determining scene semantics (Wang et al. 2006), analysis of person propagation in a single scene or across multiple video cameras (Khan et al. 2001) are based on one of two approaches, namely, (a) single person or blob analysis, implying tracking one person at a time and (b) motion analysis, implying analysis of motion fields representative of crowd or person activity. Single person analysis or the bottom up approach involves tracking individual blobs across video frames and then analysing their trajectories. Trajectory analysis forms the foundation for Brostow and Cipolla (2006) who perform unsupervised path detection obtained by clustering trajectories, Wang et al. (2006) determine scene semantics by classifying traversed paths as pedestrian walkways and vehicle pathways based on additional qualifiers such as blob velocity and blob size in addition to others. In other words at the core of surveillance analysis given the individual blob analysis is ‘the tracking problem’. A taxonomy of tracking algorithms is provided in Table 2.1 adapted from a survey paper in 2006 by Yilmaz et al. (2006).

The second approach of motion analysis, being the one observed in this thesis treats all moving objects as motion fields composed of motion vectors (speed + displacement) and clusters these for deriving models that fit real world anomalies. The approach of motion analysis for anomaly detection as opposed to blob tracking is also shared by Andrade et al. (2006) amongst others which are further discussed in Section 2.4.

Understanding the factors causing the tracking problem is key to gaining insight into the limitations of computer vision systems for purposes of complex surveillance scenes. For a feature or a blob to be successfully tracked across the footage, a number of conditions must be met, specifically:

1. The vantage point of the camera must be high if not overhead to avoid temporarily losing the object due to obstructions (**occlusion**).
2. The video quality must be high so as to successfully track an object as it moves from

the near to the far field.

3. Minimal light changes, in the absence of which a feature such as colour would change colour and no longer retain it's value as seen in previous frames.

The above conditions are typically not available in industry standard video which is generally low video resolution (approx.  $320 \times 240$  width $\times$ height) and low frame rate (typically 3 to 10 frames/sec.) with blocky compression codecs that degrade image quality resulting in degraded image features. The primary reason for broken trajectories or loss in feature correspondence is occlusion as also agreed by [Andrade et al. \(2006\)](#) and others in the research community.

Most state of the art feature trackers are known to work well with datasets containing few and well spaced people, a view shared by [Andrade et al. \(2006\)](#) and others who have adopted motion analysis for analysing crowds as they observe that even the state of the art tracking algorithms cannot deal with the problem of occlusion presented by large crowds. Tracking individuals in large crowds at high transit public locations such as bus terminals and train platforms generally results in broken trajectories or merges and splits such that there is insufficient information for reliable tracking. Thus drawing from such conditions, the single person approach impedes further processing due to broken trajectories.

Therefore, the alternative is clustering motion fields. The motion field representative of activity **over a period of time** may be viewed as a collection of completely broken trajectories which is derived by tracking a feature point between consecutive frames and is unlike person tracking which seeks to track the same feature or blob across several frames. The idea is to observe crowds as fields of motion much like a fluid but without impositions of fluid or gaseous theories upon it and simply examine/mine the motion fields over time for the presence of recurring patterns and this is precisely the approach used in this thesis.

Categories	Representative Work
<b>Point Tracking</b>	
Deterministic methods	MGE tracker ( <a href="#">Salari and Sethi 1990</a> ), GOA tracker ( <a href="#">Veenman et al. 2001</a> ).
Statistical methods	Kalman ( <a href="#">Broida and Chellappa 1986</a> ), JPDAF ( <a href="#">Bar-Shalom and Foreman 1988</a> ), PMHT ( <a href="#">Streit and Luginbuhl 1994</a> ).
<b>Kernel Tracking</b>	
Template and density based appearance models	Mean-shift ( <a href="#">Comaniciu et al. 2003</a> ), KLT ( <a href="#">Shi and Tomasi 1994</a> ), Layering ( <a href="#">Tao et al. 2002</a> ).
Multi-view appearance models	Eigentracking ( <a href="#">Black and Jepson 1998</a> ), SVM tracker ( <a href="#">Avidan 2001</a> ).
<b>Silhouette Tracking</b>	
Contour evolution	State space models ( <a href="#">Isard and Blake 1998</a> ), Variational methods ( <a href="#">Bertalmio et al. 2000</a> ), Heuristic methods ( <a href="#">Ronfard 1994</a> ).
Matching shapes	Hausdorff ( <a href="#">Huttenlocher et al. 1993</a> ), Hough transform ( <a href="#">Sato and Aggarwal 2004</a> ), Histogram ( <a href="#">Kang et al. 2004</a> ).

Table 2.1: Tracking Categories adapted from a survey paper on state of the art algorithms ([Yilmaz et al. 2006](#))

## 2.4 Related work in crowd analysis

The following pieces of work address different aspects of crowd surveillance such as crowd detection, crowd estimation/counting, determining scene semantics and detecting crowd emergencies.

[Davies et al. \(1995\)](#) experimented with a number of techniques for crowd estimation such as edge detected pixels versus the number of pre-selected background pixels and others. Their main aim was to achieve crowd velocity estimation. This was achieved by computing optic flow between image frames and then deriving velocity information from it. The authors do not use Horn and Schunck's method of optic flow ([Horn and Schunck 1980](#)) as it relates to single pixel displacement which may introduce large errors due to swaying arms, hand bags and such similar factors. They perform pixel neighbourhood (block) similarity between image frames. A  $10 \times 10$  pixel block of pre-selected foreground pixels from the first frame is compared with same sized neighbouring blocks in the second frame. Grey value distributions between the image blocks are compared using *sum of pixel-to-pixel absolute difference*. The image block from the surrounding neighbourhood from frame two with the minimum difference is selected as the region to which propagation occurred. For each tracked  $10 \times 10$  block, the motion vectors therein are aggregated resulting in a single representative vector, the reason for which is to reduce noisy vectors of the arms and legs. The dominant crowd direction is determined by adding all the derived velocity information into a polar histogram binned at an interval of  $1^\circ$ . A simple user defined threshold reveals the largest direction bin. As the authors did not implement Horn and Schunck's algorithm ([Horn and Schunck 1980](#)) for optical flow which was robust and invariant to the quantisation of brightness values as well as additive noise, the lighting conditions were kept constant.

In contrast to this early work, this thesis is able to deal with changing light conditions as a robust algorithm for optic flow computation ([Bouquet 2000](#)) is used. Furthermore, motion is analysed on a per path basis with no global thresholds for noise removal. Therefore this thesis handles a more complex task of not only learning velocities in a pathwise fashion but also the densities of crowd motion throughout the day.

[Boghossian and Velastin \(1998\)](#) created a hardware solution that operates in real-time with existing CCTV signals at 25 frames per second (fps) in CCIR/PAL mode and 30 fps in NTSC mode for analysing video for crowd emergencies and automatically inform security teams about the event. The system was specifically designed as dedicated hardware so as to render real-time results. The detection algorithm they used was computing brightness patterns across frames as devised in Velastin's previous work ([Davies et al. 1995](#)).

The main objective was hardware design that could conduct motion estimation, that is, compute direction and speed information of moving objects. Algorithmically, the authors maintained motion vectors of consecutive frames in a polar plot with some direction interval ( $\delta\theta = 1^\circ$ ) and adding the magnitudes of the vectors in each interval. Any spikes in the plot imply an emergency. The system does not maintain a history of motion (time). The hardware implementation details have been left out as this chapter concerns itself with computer vision algorithmic know-how and not the hardware implementation details. In terms of results, the system successfully performed real-time motion estimation at 25 fps video feed from a CCTV camera.

Subsequently, [Boghossian and Velastin \(1999\)](#) introduced an online system that recognised certain patterns such as bottlenecks, convergence and blockages in crowds. They used a Hough voting scheme that identified these patterns as peaks in Hough space and when detected, an alarm was signalled. This system as acknowledged by the authors needed some degree of manual tuning depending on the environment.

As mentioned in the approach in Section 1.1, this thesis uses the idea of re-enforcement to learn normal motion patterns and hence no specific abnormalities are searched for. The system does not need manual initialisation and can detect motion abnormalities as long as they significantly deviate from the normal learned definitions. Refer chapters on methodology (Chapter 3) and results (Chapter 4) chapters.

Subsequently, [Sharma \(2000\)](#) introduced a subjective factoring multi-agent system for simulating crowd behaviour in different environments. The research was aimed at predicting



crowd behaviour under extreme conditions so as to assist in designing future environments that would be prepared to avert and better deal with crowd events.

PRISIMATICA is a multi-sensor distributed surveillance framework that incorporated vision tasks such as intrusion detection in forbidden areas, detection of people going counter flow and temporal density maps for location usage (Velastin et al. 2004). The density map is simply a collector mechanism that builds up motion counts at a pixel or block level. Heavily used scene regions would collect higher motion counts than others thus showing usage differences. This thesis draws inspiration from the idea of accumulating counts and applies it by collecting features over time (feature re-enforcement) in order to discover the presence of repetition in activity. The idea of activity repetition forms the basis of building localised definitions within the scene and is explained in Chapter 3.

Brostow and Cipolla (2006) use an unsupervised Bayesian framework to track individuals in a crowd. As the features selected on a person are rigid in relation to one another and with consistent temporal tracking, resulting feature trajectories of individuals do not require further hypotheses building. Moreover, the system does not require temporal traversal and can perform reasonably without training data. However, if an individual is not detected initially, the system disregards his/her presence throughout the view of capture. Finally, the authors state that the system would not work reliably for tight cluster formations such as parades. Tight clusters of people (occlusion) would result in the loss of an individuals features and deliver unreliable motion tracks necessary for higher level processing. In contrast to this, the proposed framework would track the crowd rather than individuals, thus lost features replaced by new ones would not hinder deriving crowd-motion information for higher level processing.

Andrade et al. (2006) describe crowd behaviour as normal or abnormal by creating a framework that computes foreground optical flow, followed by principal component analysis of flow vectors, spectrally clustering these resulting in multi-object HMMs (MOHMM) class labels. Note that the classification is based on a threshold that is less than the minimum likelihood value present in the normal training dataset. If the training data proves inadequate, the

classification is likely to be incorrect. The one emergency event the system was trained for was a bottleneck event at the exit area.

The foreground motion is computed by the optical flow algorithm in (Black and Anandan 1993) and the foreground was filtered out using the foreground mask from background subtraction by Stauffer and Grimson (2000). The reduced set of feature vectors computed by applying PCA and obtaining the  $J$  highest eigenvectors is given by:

$$W_n = \{w_{n1}, \dots, w_{nT}\} \quad (2.5)$$

where,

the video segment  $V$  is divided into fixed length segments of  $T$ ,

$w_{nt}$ , a vector, represents the  $t^{th}$  frame in the  $n^{th}$  segment across the chosen eigenvectors and is given by

$$w_{nt} = \{g_{nt1}, \dots, g_{ntm}\} \quad m = 1 \dots J \quad (2.6)$$

where,

$g_{ntm}$  is the weight for the the  $m^{th}$  eigenvector.

After spectral clustering, the similarity between video segments is given by,

$$S_{ij} = \frac{1}{2} \{ \log P(W_j | B_i) + \log P(W_i | B_j) \} \quad (2.7)$$

where,  $S$  is the similarity matrix of pairwise of video segments,

$B$  represents the number of MOHMM models

Post clustering,  $K$  classes are formed and the  $W_n$  samples in each class are used to train a new MOHMM for every class  $M_k$ . The resulting model is given by,

$$P(W|M) = \sum_{k=1}^K \frac{N_k}{N} P(W|M_k) \quad (2.8)$$

where, the authors refer to  $N_k$  as the number of video segments clustered in the class  $k$  and the prior on the model weights is given by the ratio. The abnormality status of the  $n_{th}$  test segment of video,  $W_k^o$  (observation) against the bank of MOHMMs using the detection threshold is given by,

$$P(W_k^o|M) < Th_{Ab} \quad (2.9)$$

The system was reported to have correctly classified the one emergency event it was trained to detect.

The other crowd related research is crowd estimation which is an essential metric used to describe degree of safety, extent of crowd and derive average usage of an environment over time. Numerous techniques have been employed to estimate crowd density, the following are the notable ones. [Marana et al. \(1997\)](#) used texture analysis by computing a grey level density matrix GLDM that produced feature vectors representative of crowd density. These were classified into crowd-density ranges by a self-organising map (SOM) neural network.

[Casas et al. \(2005\)](#) perform face counting (frontal view) in crowded demonstrations and their system deals with occlusion, splits and crossover of humans. They use a spatial detection and temporal tracking scheme that incorporates post-processing of morphological analysis to detect skin regions of the face. As a result of this technique, non-frontal face views would go undetected leading to flawed estimates.

[Rabaud and Belongie \(2006\)](#) count people by first computing the optical flow, respawning features in motion-active regions of the image, traverse the video frames to cluster those features that have been together from inception and finally use graph partitioning to conduct agglomerative clustering. The count of the clustered feature tracks gives the count of the moving people. The system ignores by-standers as it assumes continuous human motion. Occlusion due to dense crowds would result in lossy features affecting the clustering and hence the count. The proposed research would require people estimates and would draw ideas from prior methodologies.

## 2.5 Summary

In this chapter, the tracking problem was discussed and why it remains unsuitable for crowd analysis mainly due to occlusion. The work undertaken in crowd analysis spans (a) crowd detection, (b) counting people or density estimation and (c) crowd monitoring for abnormalities, out of which extensive work is undertaken in density estimation with few undertakings in monitoring crowds using event detection or mining for normal patterns of flow. Based on the literature review conducted at the start of this thesis and recently too, work related crowd event mining use time as an extra qualification to spatial points for trajectory definition or use it as an unavoidable intrinsic feature of a video sequence. Few authors like [Andrade et al. \(2006\)](#) use time as part of their cluster definitions to make multiple models of activity. For path detection, the process is still driven by assumptions of entry and exit points, little consideration to near and far fields of the video are taken, thus removing small blobs which perhaps is due to legitimate motion in the far-field.

The next chapter explains the assumptions and ideas supporting the approach to modelling activity spatio-temporally, that is using time and space information followed by a formal notion of the technique and then system implementation steps.

## Chapter 3

# Methodology

### 3.1 Introduction

The previous chapter provided summaries of related work highlighting their strengths and weaknesses in relation to the aims of this thesis. This chapter explains the modelling technique used to achieve the thesis goals/aims. The aim of this thesis as stated earlier is to devise a system that monitors crowd activity as recorded by surveillance cameras for abnormality detection. Further, this thesis uses *motion* and *time* as the only video features to model crowd activity. These are used as explained earlier to avoid the tracking problem that limits the scope of processing.

*Motion* is defined as a flow field of motion vectors caused by a moving object and is extracted from the video using optic flow. The terms ‘public location’ and ‘scene’ are meant to imply the view of a public location or scene as seen from a video surveillance camera. The term ‘motion’ represents motion fields associated with some crowd activity and the terms ‘activity’ and ‘motion’ will be used interchangeably to refer to crowd activity.

## 3.2 Setting the scene

Public locations are known to have typical usage characteristics with respect to different regions (spaces) in the scene at different times of the day (time). That is, some regions in a scene show significant motion while others show sparse or no motion. And these *spatial* usage characteristics or activity levels in the scene might totally reverse or change gradually as the day progresses. For example, some railway platforms might be heavily used during the morning hours by work-bound crowds and not so much during the evenings as opposed to other platforms used by home-bound crowds. Alternatively, some platforms may show moderate usage throughout the day in contrast to the others. Hence, it may be reasoned that the spatial usage characteristics of a scene and the temporal changes therein are **effected** by social commitments such as work, attending school/university, shopping, visiting people, etc. As these social (weekly and weekend) commitments/events are consistent/repetitive, the spatio-temporal activity they effect at public locations across days is roughly consistent too. Consistent or similar spatio-temporal activity across days implies that the activity occurring in roughly the same scene locations within roughly the same time intervals is consistent across the days.

## 3.3 Preliminaries

Scene activity for a duration such as a day may be collectively processed or conceptualised quite literally as a 3D setting (refer Figure 3-1). The XY plane represents the scene, Z axis represents time and the mass of activity exists in the 3D space. Let the mass of activity for the given time duration and spatial extent of the scene be referred to as the body of activity. The specific thesis goal is to generate a reference body of activity that spans the duration sought to be modelled over the image plane. Such a model body of activity must contain a number of motion-variations spanning the temporal and spatial bounds acquired from a number of days and only then is it truly representative of daily activity. The model body of activity is queried by test samples in a spatio-temporal fashion, that is, in a 3D piecewise fashion for

normalcy. The 3D window has a spatial component or size parallel to the XY plane and a height along the Z axis (the temporal size). The criteria used to determine the sizes of the spatial and temporal components of the window which is used for piecewise comparisons are explained next.

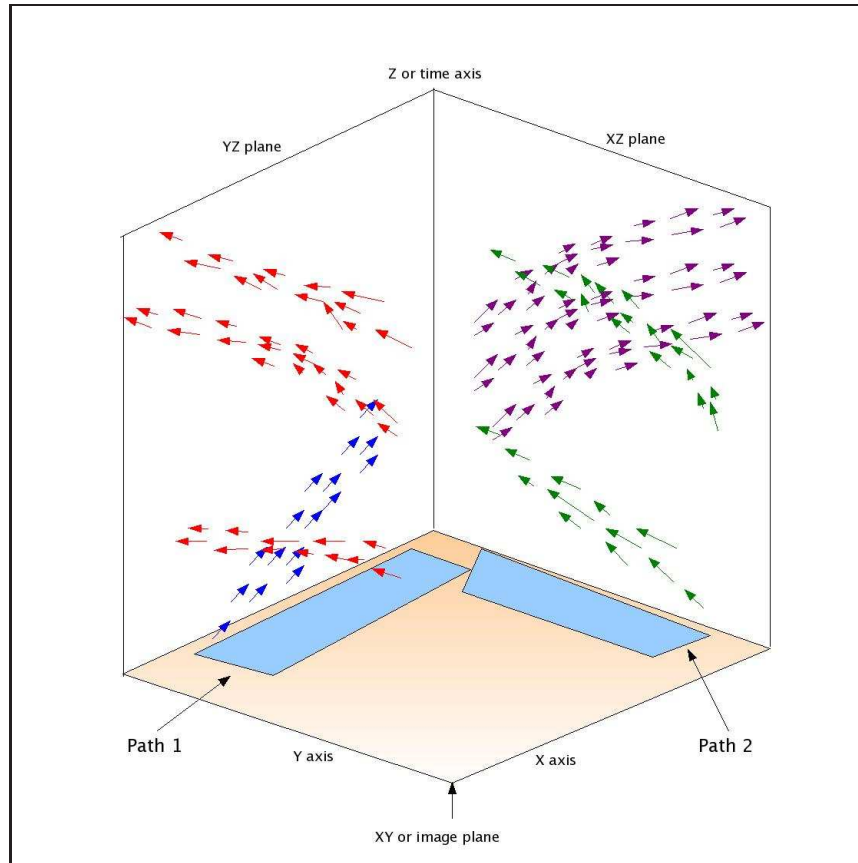


Figure 3-1: Motion along two pathways for a given period of time in a 3D setting. Z axis represents time. Motion is seen to propagate along the spatial plane (XY) and upwards along time. The different colours indicate different directions.

**Criteria for temporal component of 3D window:** Key to querying the model body of activity to assess normalcy of a test sample are the resolutions of time used for comparisons. Small resolutions (example: 2 mins, 5 mins) are too small for crowd patterns to emerge. The larger time windows of inspection bring the crowd events into focus which is ideal for comparisons.

Motion trends in a location change throughout the day and it is desirable to record these changes. However, changes in crowd patterns are seen to occur over a period of time and hence the window for inspection should be sufficiently large such that comparisons with the learned model would highlight the presence of change. Empirically this is found to be approximately 10 to 20 minutes. It takes 10 to 20 minute windows to not only notice changes in trends but also to successfully compare two intuitively similar bodies of activity (refer chapter 4 on experiments). The constituents of activity that serve as the elements for comparison are observed to be a summary of the direction flows, a summary of the speeds and a summary of the spread of the crowd over time for the given body of activity. The constituents may also be referred to as the properties of motion.

By recording the properties of motion every 10 to 20 minutes or every *time-res* minutes, one obtains a description of activity trends. By maintaining a variation of different day's motion in terms of a selected time resolution, derived is a temporal body of reference that can be used to check the normality of a test sample of activity.

**Criteria for spatial component of 3D window:** A scene is made up of differing activity regions. The meaning of differing activity regions is clarified thus. For example, waiting for a train is done on platforms and not on stairways while intra-scene propagation is observed via pathways, stairways, etc. Logical regions in a scene (example: platforms, stairways, pathways) serve a purpose and hence have a limited number of motion patterns that intuitively define it. The phrase 'limited number of patterns' implies a few patterns repeating themselves over time but in different quantities. For example, (a) a railway platform shows intuitive patterns such as people waiting, entering trains, departing trains and passing by along the length of the platform, (b) a pathway usually exhibits two-way motion, the densities of which are seen to change over time. Therefore these regions acquire their logical name not just by one or two patterns in a **small** time span but by the re-enforcement of the occurring patterns captured over a larger time span. Regions showing a temporally homogeneous mix of repetitive motion patterns are referred to as *activity regions* and such regions are learnt by growing unit areas or regions based on the similarity of their *collective temporal motion patterns*. The above concept of activity recording is meant to be applied using space as well as time. The spatial



component of choice is the activity region. In the body of activity, smaller entities are defined spatially by activity regions and temporally by a suitably selected temporal resolution. Such a scheme divides the larger body of activity into logical chunks that serve as the means for space-time queries.

Recording activity over scene regions based on any other scheme for the spatial component would result in a mix of dissimilar motion patterns having unrelated occurrence contexts. It is changing trends of activity that is the requirement for recording and this carries the silent assumption that the activity is coherent and inherently has patterns (repetition with variations) and therefore activity recording is performed over motion similar regions. The terms *paths* or *activity regions* will be used interchangeably to refer to motion similar regions.

**A note on paths:** The conventional meaning of paths in a scene refers to the observed paths. Some scenes might show pathways starting in the far-field and terminating in the near-field. The definition of activity regions would in fact segment such an observed path into a number of motion similar regions primarily due to changing speeds of motion caused by near/far-field phenomenon. Merging these regions over a single path would require scene-motion normalisation which in turn requires depth of scene information. As this system is intended to be used generically, depth of scene information is unavailable and hence any normalisation factor is unavailable too. However, the purpose for the normalising factor would be to normalise motion across a path so as to perform comparisons across the path in order to detect abnormalities, etc. Even though the single long path may be segmented into activity regions, each activity region would be qualitatively more localised in recording motion as opposed to the whole path and any abnormalities along the path ought to be detected by the local models. In this thesis, the term path is meant to imply activity regions which may not necessarily be the observed paths.

From the foregoing, activity in a scene may be recorded in manageable chunks obtained by dividing the 3D setting of motion (space x,y and time z) into vertical columns based on the learnt XY paths and then each column is divided into discrete fixed length durations based on the empirically derived time resolution.

In order to easily represent the millions of motion vectors that make up the body of activity in the 3D setting, a compressed summary of motion is required. As discussed earlier the properties of interest of a body of motion are the number of direction flows, the speeds observed and how the different flows are distributed over time (spread). Two intuitively similar bodies of motion imply that they contain roughly the same number of direction flows (example: 2 way motion) and that each of the motion bodies have roughly the same spread over time with roughly the same speeds. Two intuitively similar entities/bodies of motion may exhibit any number of variations in their properties that would not compare as similar if compared using direct time coincidence. The matching scheme used must be **invariant** to the variations of the properties but capture the intuitive similarity. In order to capture the intuitive similarity, the storage or recording mechanism of the motion is targeted and not the comparison technique. The properties of interest of a motion entity that sufficiently define it have been observed to be (a) the observed directions, (b) observed speeds, (c) and the observed spread of the motion over time. The storage representation for each of these is selected as *a one dimensional histogram*. The direction histogram implicitly captures the quantities of motion vectors for the different direction flows. The speed histogram presents the distribution of speeds observed and the spread histogram captures the intuitive spread of the motion over time. For the spread histogram each bin represents the counts of motion vectors/unit time and the number of bins completes the time sample. The bins for the direction and speed histograms are selected appropriately so as to capture as many distinct flows and speeds as possible and is quantified in the next chapter.

The advantages of using histograms for recording motion are as follows:

1. It is invariant to the natural variations of motion and implicitly captures the descriptions of interest, namely (a) quantities of motion in different directions (density of flows along different directions), (b) overall spread of motion. It is relatively less important how the directions are scattered across the observed spread as long as their summed quantities are consistent across two entities of motion.
2. Any body of motion contains noise, that is motion vectors due to flickering light, swaying hands, swinging legs, etc. where the vectors of interest are those of the torso. It is

not possible to learn the noisy directions as people may walk in any combinations of meandering routes over a path, however, the noisy vectors were found to be consistent in the directional histogram bins. Therefore *factoring* the noise into the data is justifiable. Thresholding the noisy bins is risky as the supposedly small counts of noise might in fact be a legitimate sparse flow.

3. Finally, the most significant reason for recording **observed** directions in histograms is that the observed distribution of motion are merely **recorded** (in discrete bins) as opposed to (say) an alternate scheme that attempts to find 3D flow volumes so as to acquire explicit knowledge of the number of distinct flows. Attempting to search/segment motion flows across a path implies in the matching thresholds the silent assumption/imposition that there are well defined flows along the path. Such a technique would produce impressive results in datasets showing well defined packed flows of motion, however, datasets such as railway platforms exhibiting dispersive motion as well as well defined flows would challenge the alternative scheme. The storage structure of choice, the histogram captures any distribution of motion (dispersive or well defined) and therefore makes the system scalable for any class of motion as long as it presents itself consistently over time.

Summary of activity recording for a scene over time:

- An entity of motion is described using three descriptors, namely, direction, speed and spread. Each of these assume a one dimensional histogram. Each bin of the spread histogram represents the number of motion vectors/unit time and the number of bins completes the time sample. The speed and direction histograms are binned so as to capture a sufficient range of speeds and directions.
- In order to record activity logically and in manageable chunks, the body of activity is first divided into vertical columns based on the number of paths (activity regions) along the XY plane and then each column is divided into fixed length durations, the duration of which is the time resolution derived empirically. Thus activity is segmented along space based on activity similarity and along time using a resolution that facilitates

not only detection of changing trends but also comparisons between intuitively similar activity.

- Finally, a single space-time chunk resulting from the above division summarises it's contained activity via the scheme of histograms. Each space-time chunk is referred to as *a voxel of activity* and the resulting space-time divided 3D system is referred to as *a system of voxels*.

The next sections provide the formal notations, conceptual design and equations for generating a model system of activity and using it for comparisons.

### 3.4 System Overview

Figure 3-2 summaries the system overview. Central to generating a model and testing unseen data is the 3D system of voxels that facilitates querying data as per path indices and time (time resolution steps  $t$ ). Step A in the figure represents the process of learning the number of paths and deriving the appropriate time resolution. The outputs of Step A are the learned path labels and time labels that may be used to label training data for the model generation process or test data for sample classification process. The output of Step A is not any real usable motion data but simply the structure needed to facilitate queries and comparisons.

Step B in the figure represents the generation of the model. In this step, the inputs are training data sets labelled using learned spatio-temporal labels obtained from Step A. The output of this step is a trained model containing variations of direction, speed and spread for every voxel in the system. Step C represents the steps to perform classification of a test sample where the test data is labelled using spatio-temporal labels and then compared against the model generated from Step B. The following sections describe in detail the steps used during system implementation.

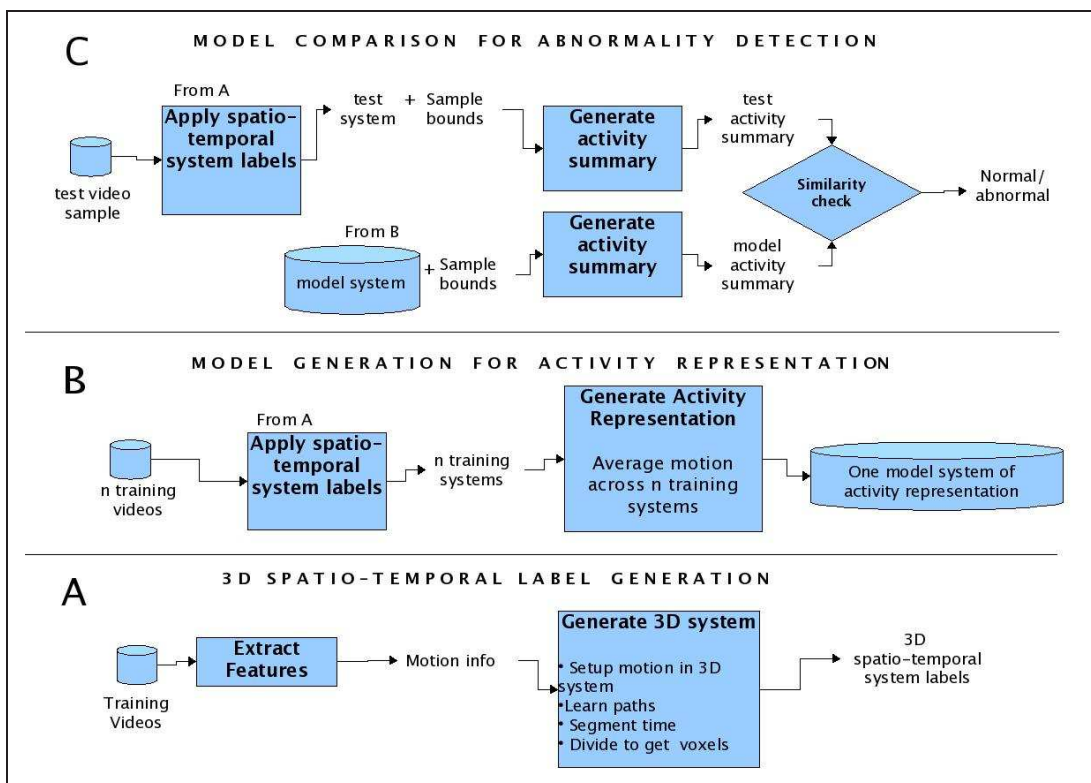


Figure 3-2: System Overview diagram outlining the approach used to generate activity representations (Step B) which is used in Step C for abnormality detection. Both Steps B and C depend on A where path and time labels (spatio-temporal labels) are generated.

### 3.5 Feature Extraction

The input for this step is a video of a scene for a given duration and the output is a textfile of time indexed (video frame numbers) foreground motion information. The following sequence of processing is performed to obtain foreground motion:

- Optic flow ([Bouquet 2000](#)) is computed between every consecutive video frame. The computation uses a large number of optic flow features per frame which sufficiently account for even dense as well as sparse foreground motion.
- Prior to computing the optic flow, median smoothing is performed across frames using a 3x3 kernel. This is done in order to stabilise non-moving regions (background) in the scene so as to reduce noise or image jitter for the optic flow computation.
- Following computation of optic flow, background subtraction ([Stauffer and Grimson 2000](#)) is computed between the same two frames and the resulting foreground mask is used to obtain foreground optic flow/motion. The adaptation rate, the number of Gaussians in the model and the standard deviation of the background Gaussians measures are provided in the next chapter.
- The motion information derived from this exercise are (a) the start and end coordinates of each tracked feature point, that is,  $(x_1, y_1)$  &  $(x_2, y_2)$ , (b) the frame number (time) for each foreground motion vector. These quantities are then written to a text file. The process iterates until all the frames for the given duration have been processed. [Table 3.1](#) is an example of the text file where every row represents a single motion vector for a given frame. A video with typical activity that is one hour long and processed at approx. 8 fps outputs approximately 700,000 motion vectors.

**Noise thresholding:** While computing the optical flow, no thresholds are used to filter out spurious flow vectors. This is primarily because the scale of the data and depth of scene information is unknown. Once the motion information is extracted, the noisy motion vectors are removed. When performing this, no assumption on the direction component of

x1	y1	frame_number	x2	y2
121	5	30000	121	4
212	12	30001	212	11
286	26	30001	286	26
288	26	30001	288	26
289	30	30001	289	30
287	21	30001	287	21
122	11	30001	122	11
125	8	30002	125	8
286	31	30002	286	31
290	31	30002	290	31

Table 3.1: Example of a text file containing motion information

a motion vector is made as noisy directions may infact be legitimate sparse flows. Hence the displacement component of the vector is targeted. Only very large displacement vectors are removed from the dataset. These would lie outside 3.5 standard deviations of the mean displacement of the dataset.

### 3.6 3D Spatio-temporal Label Generation

The spatio-temporal system is simply two sets of labels, one for the spatial activity regions/paths and the other for segmenting clock time into intervals of length  $t$  (temporal resolution). The path labels are learned from the data while the magnitude of  $t$  is obtained empirically (refer chapter 4). Path labels may be derived only after motion information is extracted from the training videos.

The model to be generated is generated so in order to model a given duration of the day, for example 5 hours, 2 hours, a whole day, etc., therefore *a single set* of training data implies  $n$  hours of video (example:  $n=5$  hours) and  $m$  training data *sets* implies,  $n$  hours of daily video for  $m$  days. The training data for learning spatio-temporal labels is a subset of the  $m$

training sets.

### 3.6.1 (i) Generating/Learning Path Labels

Section 3.3 describes activity regions as regions whose temporal collection of occurring patterns are similar. That is, regions are defined so if the area they bound presents similar recurring patterns of motion. This definition dictates the scheme used for learning activity regions or paths.

**Summary of the process:** Motion data extracted from the video is setup in a 3D setting such that the XY plane represents the image plane, Z represents time and the motion data exists in the 3D space. In order to identify unit areas in the scene as having recurring patterns, the time coordinate for each motion vector is ignored resulting in a projection (literally) of all motion vectors upon the XY plane. Thus a rich homogeneous mixture of motion regions are created that naturally define their own boundaries. The task now is to devise a technique to capture the mix of motion (direction+speed) for comparisons that would facilitate region growing. This is done by setting up a grid over the XY plane and summarising the motion in each grid block with a direction and a speed histogram. The term *motion similar regions* implies that each region has a similar mix of directions and that their displacement sizes are similar too. Motion regions are derived by first connecting grid blocks based on similar direction, then maintaining a separate record of connecting neighbours based on speed and finally performing an intersection of speed and direction labels such that each direction label has only one speed label.

#### **Implementation of Path Learning:**

- Motion information as described in Table 3.1 is setup in a suitable complex 3D data structure. The direction (in degrees) and speed (in pixels) for each motion vector is computed.
- A grid as explained above is setup to lie above the image plane and divide it into  $M$



$\times N(\text{width} \times \text{height})$  blocks. Refer chapter 4 for values of  $M$  and  $N$  which were observed to capture unit area similarities well with results closest to the ground truths for two public location videos.

- Motion in each grid block is summarised using a one dimensional direction histogram and a one dimensional speed histogram. The direction histograms span the range from  $-180^\circ$  to  $+180^\circ$  and are binned so as to capture direction groups sufficiently. The speed histograms are binned over the observed range of displacement values. The binning scheme for the speed histogram showed more accurate results in two public scene datasets (refer chapter on Experiments, chapter 4) while the direction histogram was binned thus to capture sufficient detail of motion flows. Larger number of bins removes the generality of motion flows by highlighting the meandering flows of people along a straight path while few bins (example: 4 or 8) imposes the assumption that the paths in the scene are aligned with the binning scheme as practised in (Wang et al. 2006).
- Two iterations of region labelling are performed over the grid where in each iteration, each grid block is compared with its eight neighbours. The first iteration is based on direction similarity while the second iteration is based on speed similarity. Here a label refers to a region (collection of neighbouring blocks). If the scene semantics are subject to the near and far field phenomenon, then at least one direction label may have more than one speed (displacement size) label or vice-versa. Thus the final step is to divide either speed labels or direction labels such that the resulting regions of blocks contain a unique direction label and a unique speed label.

The above steps completes the process of path (activity region) learning. The above approach of using direction and speed grids is similar to the process followed in (Zhan et al. 2005), however, the authors attempt to derive a single representative direction and speed measure for each grid block that forces the assumption of one-way motion along a detected path. In contrast, this thesis retains the accumulated direction and speed distributions thus accounting for any velocity mixture that may occur.

### 3.6.2 (ii) Generating Temporal Labels:

The motion data is segmented into durations of size  $t$ . The value of  $t$  could be approximately 10 to 20 minutes as this fits well with clock time, is large enough to allow pattern matching and small enough to suppress fine grain dissimilarities.

This step involves simply labelling each motion vector in terms of  $t$  intervals such that the first  $t$  minutes of motion vectors are labelled 1, the next  $t$  minutes of motion vectors are labelled 2, etc. till  $T$  windows each of size  $t$  are completed.

This completes the step for learning spatio-temporal labels. The spatio-temporal labels represent the phase for learning where in the scene the paths lie and together with the temporal labels, they act as an indexing scheme to query a part of the motion mass in a 3D system. The model generation phase which is next uses the spatio-temporal labels to access motion in a structured fashion at a given time and space for generating activity representations in that time and space indexed as  $(p, t)$ , that is  $(pathnumber, timeindex)$ . The spatio-temporal labels provides the voxel definition/structure as seen in Figure 3-3.

## 3.7 Model Generation for activity representation

The spatio-temporal system in which motion data is segmented into space and time is visualised in fig. 3-3. The XY plane represents the image plane containing paths, time is along the Z axis and the 3D space is where the motion is seen to occur moving along the XY plane and naturally upwards along time. The 3D system is seen to be divided spatially as per the number of paths and temporally as per the empirically derived time resolution. Each time-space division is called a *voxel* of activity and such a 3D setting is called a *system of voxels* or simply *system*. Let the system be denoted by  $\lambda(p, 1 : T)$  where  $p$  is the path index and  $1 : T$  is simply the range of time resolution intervals from 1 to  $T$ . The activity of a given voxel bounded by the underlying path  $p$  (space) and temporal resolution as height starting

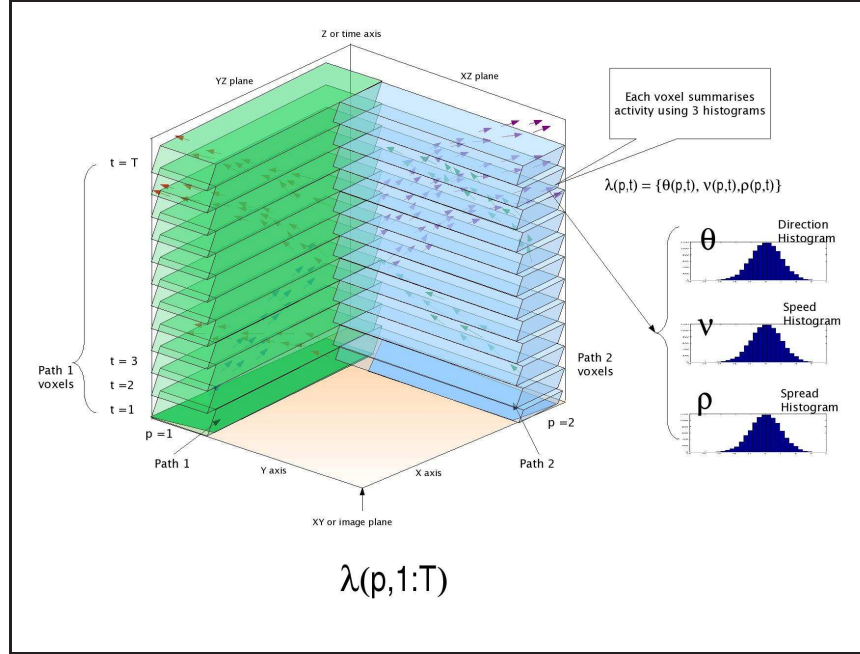


Figure 3-3: 3D system of voxels  $\lambda$ . Activity representation for a single day or time period. Each voxel contains 3 histograms - direction ( $\theta$ ), speed ( $\nu$ ) and density ( $\rho$ ) describing the motion in that space and time.

at a given time  $t$  is summarised as three histograms, namely direction, speed and spread as follows.

$$\lambda(p, t) = \{\theta(p, t), \nu(p, t), \rho(p, t)\} \quad (3.1)$$

such that.

$$\begin{aligned} \theta_i(p, t), \quad i &\in \{1, 2, 3, \dots, B_\theta\} \\ \nu_j(p, t), \quad j &\in \{1, 2, 3, \dots, B_\nu\} \\ \rho_k(p, t), \quad k &\in \{1, 2, 3, \dots, B_\rho\} \end{aligned}$$

where,  $B_\theta, B_\nu$  and  $B_\rho$  are the number of bins in direction ( $\theta$ ), speed ( $\nu$ ) and spread ( $\rho$ ) histograms.

- $\theta$  is divided into a range of bins so as to capture a large range of directions in degrees.
- $\nu$  is divided into a number of bins to capture the displacement of motion vectors in pixels.

- $\rho$  is generated such that each bin represents the number of motion vectors observed in unit time (density) and the number of bins completes the duration of the temporal resolution  $t$ .
- All histograms are normalised over one. (example:  $\sum_i \theta_i = 1$ ).

The distinction between training, model and test data is as follows: (a) model data is superscripted with an  $M$ , (b) training data is superscripted with  $*$  and (c) test data is superscripted with a  $T$ .

As described in the section 3.3, the model is much like the current system of voxels but with variations of activity obtained from different days across the same time and space. The model system of voxels therefore has vectors of direction, speed and spread histograms gathered from different days of the week (refer fig. 3-4). The model system is represented as  $\Lambda(p, 1 : T)$ , where as before  $p$  provides the path index and  $1 : T$  is the range of temporal resolution intervals from 1 to  $T$ . Activity for a single voxel of path  $p$  and temporal resolution  $t$  is given by:

$$\Lambda^M(p, t) = \{\Theta^M(p, t), N^M(p, t), R^M(p, t)\} \quad (3.2)$$

such that,

$$\begin{aligned} \Theta_i^M(p, t), i &\in \{1, 2, 3, \dots, n_\Theta\} \\ N_j^M(p, t), j &\in \{1, 2, 3, \dots, n_N\} \\ R_k^M(p, t), k &\in \{1, 2, 3, \dots, n_R\} \end{aligned}$$

where,

$\Lambda^M$  is the model of activity representation by way of a system of voxels.  $\Theta$ ,  $N$  and  $R$  are direction, speed and spread vectors respectively. Each vector contains a number of histograms obtained from training data which will be explained next, hence the capital letter notations.

Each voxel in the model system represents activity for the same time-space ( $p, t$ ) over

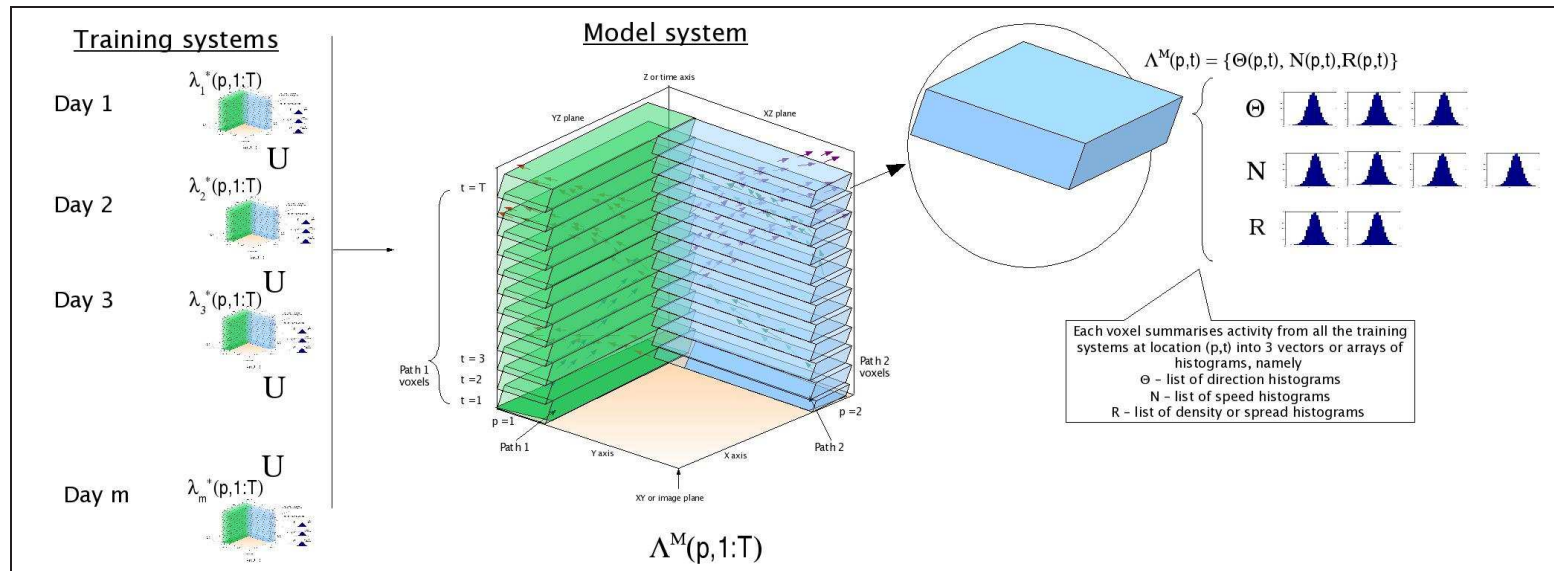


Figure 3-4: Model system of voxels represented by  $\Lambda^M$ . Each voxel above a given path contains a 3 vectors of histograms - direction ( $\Theta$ ), speed ( $N$ ) and density ( $R$ ) aggregated from different days over the same space and time.

several days. This is achieved as follows:

The body of activity spans a given duration example: 5 hours, day, etc. The training data is therefore video data of the given duration over a number of days . Let the given duration over a single day be referred to as a set, then each set of training data is setup into a 3D spatio-temporal system. That is,  $m$  sets training data are setup in  $m$  systems. This is expressed as:  $\lambda_1^*(p, 1 : T), \lambda_2^*(p, 1 : T), \lambda_3^*(p, 1 : T), \dots, \lambda_m^*(p, 1 : T)$ . The model is generated by merging all the training systems via a union operation.

$$\Lambda^M(p, 1 : T) = \{\lambda_1^*(p, 1 : T) \cup \lambda_2^*(p, 1 : T) \cup \lambda_3^*(p, 1 : T) \cup \dots \cup \lambda_m^*(p, 1 : T)\} \quad (3.3)$$

The union operation for a single voxel given by  $(p, t)$  in the model system is as follows:

$$\Lambda^M(p, t) = \{\lambda_1^*(p, t) \cup \lambda_2^*(p, t) \cup \lambda_3^*(p, t) \cup \dots \cup \lambda_m^*(p, t)\} \quad (3.4)$$

The union operation for the above voxel is simplified in terms of voxel components as under:

$$\Theta^M(p, t) = \{\theta^{\lambda_1^*}(p, t) \cup \theta^{\lambda_2^*}(p, t) \cup \theta^{\lambda_3^*}(p, t) \cup \dots \cup \theta^{\lambda_m^*}(p, t)\} \quad (3.5)$$

$$N^M(p, t) = \{\nu^{\lambda_1^*}(p, t) \cup \nu^{\lambda_2^*}(p, t) \cup \nu^{\lambda_3^*}(p, t) \cup \dots \cup \nu^{\lambda_m^*}(p, t)\} \quad (3.6)$$

$$R^M(p, t) = \{\rho^{\lambda_1^*}(p, t) \cup \rho^{\lambda_2^*}(p, t) \cup \rho^{\lambda_3^*}(p, t) \cup \dots \cup \rho^{\lambda_m^*}(p, t)\} \quad (3.7)$$

Here,  $\Theta_M$ ,  $N_M$  and  $R_M$  are vectors of histograms containing the collection of  $\theta$ ,  $\nu$  and  $\rho$  obtained from the same  $(p, t)$  across the  $m$  training systems. After the above three union operations, the contents of each vector are  $n_\Theta$ ,  $n_N$  and  $n_R$  dissimilar histograms respectively. The similar histograms are found to be so by using normalised cross correlation with a high correlation threshold (example: 0.95) and these are merged by averaging while the dissimilar ones are maintained for variation. Hence the use of the phrase dissimilar histograms. By merging the similar histograms, efficiency of the system is increased thus addressing the concern of redundant or duplicate histograms.

$$\Theta_i^M(p, t), i \in \{1, 2, 3, \dots, n_\Theta\}$$

$$N_j^M(p, t), j \in \{1, 2, 3, \dots, n_N\}$$

$$R_k^M(p, t), k \in \{1, 2, 3, \dots, n_R\}$$

The model system denoted by  $\Lambda^M$  is generated by performing a union operation over the direction, speed and density histograms independently in each of the voxels, the collection of which (histograms) are obtained across the training systems. The union operation simply means merging similar histograms and representing them as a single averaged histogram while the dissimilar histograms are maintained as they are, thus  $\Theta^M$  results in a collection of dissimilar histograms. Similarly the speed and spread vectors,  $N_M$  and  $R_M$  respectively are made more efficient.

### 3.8 Model Comparison

Given the generated model a test sample may be compared with it in a time-space fashion to determine it's normality. Lets refer to the term sample as motion belonging to a given duration of time and a given path. The test video from which the test sample is selected is first setup in a 3D system of voxels and is denoted as  $\lambda^T(p, 1 : T)$ . The test sample denoted as  $\lambda^T(p, t_a : t_b)$  is then compared with the model quantitatively based on the number of temporal resolutions  $t$  that complete the duration from  $t_a$  to  $t_b$ . The comparison between the sample and the model over a single  $t$  is as follows:

$$\text{compare}(\lambda^T(p, t), \Lambda^M(p, t))$$

where,

the *compare()* function compares  $\theta$ ,  $\nu$  and  $\rho$  from  $\lambda^T(p, t)$  respectively against vectors  $\Theta$ ,  $N$  and  $R$  from  $\Lambda^M(p, t)$  in order to find at least one match for each activity descriptor. The individual steps of the *compare()* function are given below:

$$\delta_\theta = \text{corr}(\theta^T(p, t), \Theta^M(p, t)) > \tau_\theta \quad (3.8)$$

$$\delta_\nu = \text{corr}(\nu^T(p, t), N^M(p, t)) > \tau_\nu \quad (3.9)$$

$$\delta_\rho = \text{corr}(\rho^T(p, t), R^M(p, t)) > \tau_\rho \quad (3.10)$$

where,

$\theta^T$  is cross correlated against each  $\theta$  in  $\Theta^M$  and if even one match is found, then  $\delta_\theta$  equals one (normal), else  $\delta_\theta$  equals zero (abnormal). Similarly the values for  $\delta_\nu$  and  $\delta_\rho$  are computed.

Activity in the sample voxel indexed by  $(p, t)$  is classified as normal or abnormal by the heuristic in algorithm 1. The case of partially abnormal is avoided as this would make the classification fuzzy which is not desirable without conclusive and consistent evidence.

---

**Algorithm 1:** *Heuristics to classify outcome of summary comparison as normal or abnormal.*

---

**Input:**  $\delta_\theta, \delta_\nu, \delta_\rho$

**Output:** normality status

**if**  $(\delta_\theta \neq 1) \quad ||| \quad (\delta_\nu \neq 1)$ **then**

|  $status \leftarrow abnormal;$

**else if**  $(\delta_\rho \neq 1)$ **then**

|  $status \leftarrow partially\ abnormal;$

**else**

|  $status \leftarrow normal;$

**end**

---

- A sample is compared with the model to determine deviation from the model. Intuitively, if the quantities of direction flows in a sample appear reversed with respect to the model or new mix of direction flows (previously unseen in that space and time) appear, then this is considered to be the most alarming factor for deviation from the model and hence  $\delta == 0$  is reason enough for the sample to be assessed as abnormal.
- The variation in speeds for a voxel given by  $(p, t)$  are recorded as different histograms. The training data is so chosen so as to contain examples of occasional changes in speed with their associated quantities of motion vectors. For example, sometimes small groups of school children may be seen to run across a pathway or an occasional person running by to catch a bus or train; such examples are factored in implicitly into the higher displacement bins  $j$  of the training speed histogram  $\nu^*$ . Therefore, test samples containing a few instances of people running would show similar bin counts as those from the training samples and be classified as normal. For  $\nu^T$  to be abnormal, a significant change in bin counts due to large or small displacements is required implying a macro event (large crowd) showing abnormal consistency in the large or small motion vectors.
- The spread of data provides an intuitive insight into the patterns of flow. Markedly



different scatterings of flow are of interest. For example, if the same number of direction flows occur (a) in large chunks, or (b) thin strip of motion and if these differ from say a routine homogeneous mix, then such changes in scatterings are of importance as it implies some change in higher level events even if the quantities of directions and their associated speeds adhere to the learned model. Hence deviations in spread are not altogether abnormal, therefore termed as partially abnormal.

### 3.9 Summary

The chapter defines a formal scheme for decomposing activity into three descriptors so as to create a means for comparisons of crowd activity. More specifically, a model assuming an intuitive setting containing variations of activity descriptors for each time-space is generated. Using such a model, a test sample may be compared against the model in a spatio-temporal fashion and be quantitatively analysed for similarity, that is, the sample duration is compared with the model in a piecewise fashion based on  $t$  (temporal resolution) steps.

The next chapter describes the datasets, experiments to validate the approach towards activity modelling and settings used in the implementation.

## Chapter 4

# Datasets & Experiments

### 4.1 Introduction

The previous chapter explained the modelling and implementation of the framework used for crowd modelling and abnormality detection. This chapter explains the experiments conducted to evaluate and validate the proposed ideas for activity modelling and discuss the datasets used for testing the system. The chapter is organised as follows: an overview of the experiments is provided after which are sections describing datasets used and results of the experiments. The section on datasets explains the datasets including the abnormal activity samples used for testing the system. The experiment's section lists the three experiments used to validate the system. For each experiment is provided (a) a brief description of the experiment, (b) method of ground truthing, (c) parameters used in the implementation, (d) details of video samples used and (e) results of the experiments. The chapter concludes with the summary.

The experiments have been devised to test the techniques used in this thesis to model activity and the performance of the proposed system. The proposed technique of modelling activity using direction, speed and spread is tested on it's ability to decompose the activity and will be verified in experiment 2. The paths which provide the spatial component for queries is tested in experiment 1 (core component of the system). Finally, experiment 3 tests the system as a whole.

The first experiment verifies path learning to be consistent which validates it's process for reliability. As the first dataset (university) is limited, only one sample is used from this dataset.

The second experiment verifies two aspects: (a) that the proposed idea of modelling activity using three histograms works, (b) the time resolution empirically derived really does reveal crowd descriptions in a way that two similar samples of activity are indeed found to be similar and two dissimilar samples are found to be dissimilar. The similar and dissimilar samples are of the same temporal length but need not coincide with clock-time.

The third experiment tests the system as a whole attempting to verify that the trained model classifies normal and abnormal samples correctly. The significance of this experiment is not just testing the system but also to validate the method used for maintaining variations of activity that is meant for rendering the model as representative of daily activity. That is, each voxel of the model system  $\Lambda_M(p, 1 : T)$  contains vectors for direction ( $\Theta$ ), speed ( $N$ ) and spread ( $R$ ) that store the descriptors of the same  $(p, t)$  gathered from different days and this proposed scheme of variation ought to sufficiently classify daily activity appropriately.

## 4.2 Datasets

Surveillance videos from two public locations will be used to assess the system. They are: (a) Curtin university pathways, (b) a busy passageway of Perth train station. The two datasets are selectively used for the three experiments.

Experiment 1 aims at showing that the path detection technique produces consistent results. The university dataset is only approximately an hour long and hence is the only sample for testing for that dataset. However, it has a relatively more complex structure of observed paths than the second dataset. Both the datasets are used for path detection and this is for two reasons, (1) the second dataset (railway passage) shows a single passage which is simplistic for path segmentation but would show a couple of segmentations along the single observed

path based on the notion of activity regions.(2) The university dataset has three paths, one of which starts in the far-field and terminates in the near-field. It is of interest to determine how observed path boundaries separate from each other based on the notion of activity regions.

In experiment 2, the concept of activity modelling is verified by testing two similar and dissimilar samples. The pairs of samples are selected from different paths of the university and railway passage datasets.

Finally, as the university dataset is only about an hour long (insufficient training data) and sequences of abnormalities were unavailable, only the railway passage dataset is used for experiment 3 which deals with training the model and using it to classify test samples.

#### 4.2.1 Dataset 1: Curtin university pathways

The scene in Figure 4-1 shows three paths of a university. The busport (cannot be seen) is in the far-field just beyond path 3. Paths 1 and 2 start at path 3 and provide passage into the university. Path 1 starts in the far field and terminates in the near-field, while paths 2 and 3 lie mainly in the far-field. Figures 4-2(a) and 4-2(b) are examples of sparse and dense motion along these pathways.

The video for this location is approximately an hour long captured from 10am on a weekday. It shows crowds of people entering the university along paths 1 and 2. Each time a bus arrives at the university busport, a burst of crowds are seen passing through these paths. Path 1 shows higher usage than path 2. Occasionally people are seen leaving the university along paths 1 and 2. Path 3 facilitates activity along the perimeter of the university.

The video (colour) resolution is  $320 \times 240$  (w×h), recorded at 25 fps but processed at 6 to 8 fps for the optic flow and background subtraction. The lower frame-rate is closer to industry conditions that range between 3 and 10 fps. Figures 4-3(a) and 4-3(b) show 3D views of the raw data binned into 8 directions for distinction.

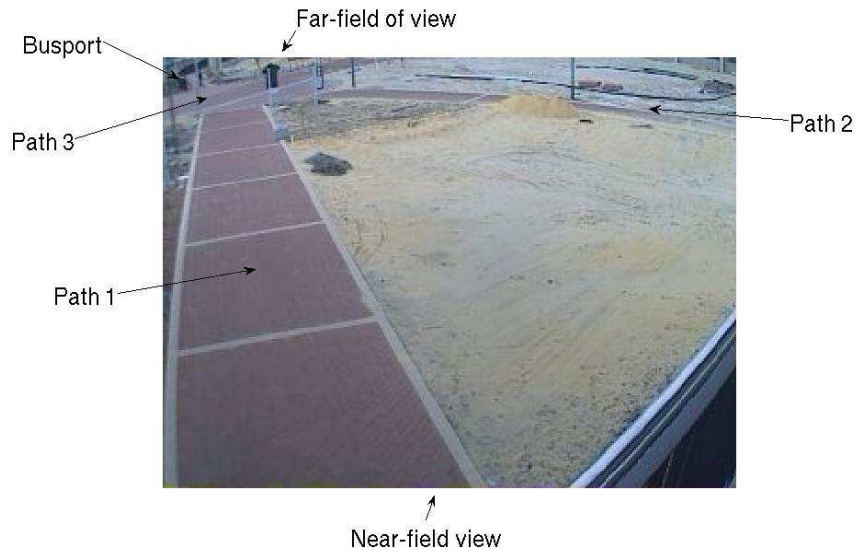


Figure 4-1: Empty scene of pathways at a University. Observed paths 2 and 3 lie in the far-field while path 1 extends from the far to the near-field of the camera view



(a) Sparse crowds

(b) Dense crowds

Figure 4-2: Examples of Sparse and Dense usage of pathways at the University.

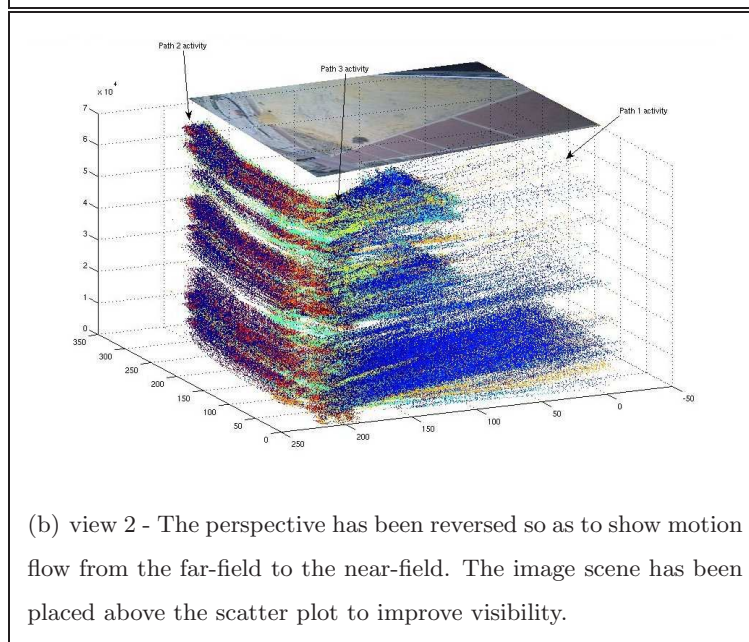
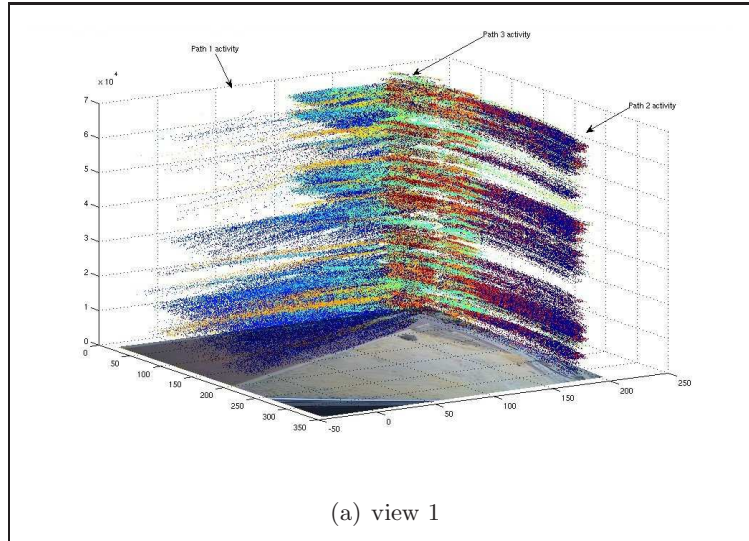


Figure 4-3: 3D scatter plots of activity spread over 60 minutes ( $Z$  axis) across the scene. The different colours indicate different direction bins. Each path presents roughly two colours above it indicating two-way motion.

The motion above each path is somewhat divided into two colours indicating two way motion. The colour showing more density and quantity represents inflow of students into the university and the colour having few occurrences of flows is representative of the outflow of students. As can be seen the trails of motion along a path are delimited by empty spaces showing the bursts of crowd motion effected by incoming busses at the busport. Few bins have been used to represent the motion flows so as to suggest the general flow of motion and acquire an intuitive feel of activity spread over time.

#### 4.2.2 Dataset 2: Perth train station passageway



(a) Empty scene

(b) Sparse crowds



(c) Dense crowds

Figure 4-4: Sparse and dense usage of railway passageway.

The passage way in the scene connects Perth underground train station and the above ground train station. Crowds of people leave the near-field of the scene to make their journey to the outdoor or train platforms above while people walk along the passage starting at the near-field

and exit the far-field in order to go to the underground train platforms. Figure 4-4 shows examples of empty, sparse and dense crowds across the scene.

The data is collected over six hours between 8am and 9am and 10am to 11am across Tuesday, Wednesday and Thursday and one additional hour of footage for Wednesday 10am to 11am taken from another week in which abnormal acts were staged. From the collection of video, two weekdays (Tuesday and Thursday) are used to train the model so as to classify the third day of footage (Wednesday). Meaning, Tuesday symbolically represents the first half of the week and Thursday represents the second half of the week (not weekend), therefore these two days are used to train the model so as to classify the middle of the week, that is Wednesday.

The video was processed in colour at 6 fps at resolution  $240 \times 180$  (w×h) for optic flow and background subtraction. From the list in Table 4.1, video of Tuesday and Thursday were used as training data in order to classify test data Wednesday. Wednesday data contains normal and abnormal videos sequences. The abnormal sequence is present in a video from 10am to 11am.

Day	Duration		Total	Type
Tue	1 × 8am-9am	1 × 10am-11am	2 hours	train
Wed	1 × 8am-9am	2 × 10am-11am	3 hours	test
Thu	1 × 8am-9am	1 × 10am-11am	2 hours	train

Table 4.1: List of videos samples used of the train station passage

Figures 4-5(a) and 4-5(b) are 3D activity plots of typical activity observed in the scene from 8am-9am and 10am-11am on week days. The plot has been oriented to show the length of passage with the near-field located at the right and the far-field at the left. As such two colours are seen out of which the majority represents crowds moving from the far-field to the near-field, that is the majority of motion is seen to move from the underground platforms to the external ones through the passageway while only few move along the opposite direction from the external platforms to the underground ones. Visual inspection shows a slight decrease or sparseness in activity between 8am-9am and 10am-11am data.



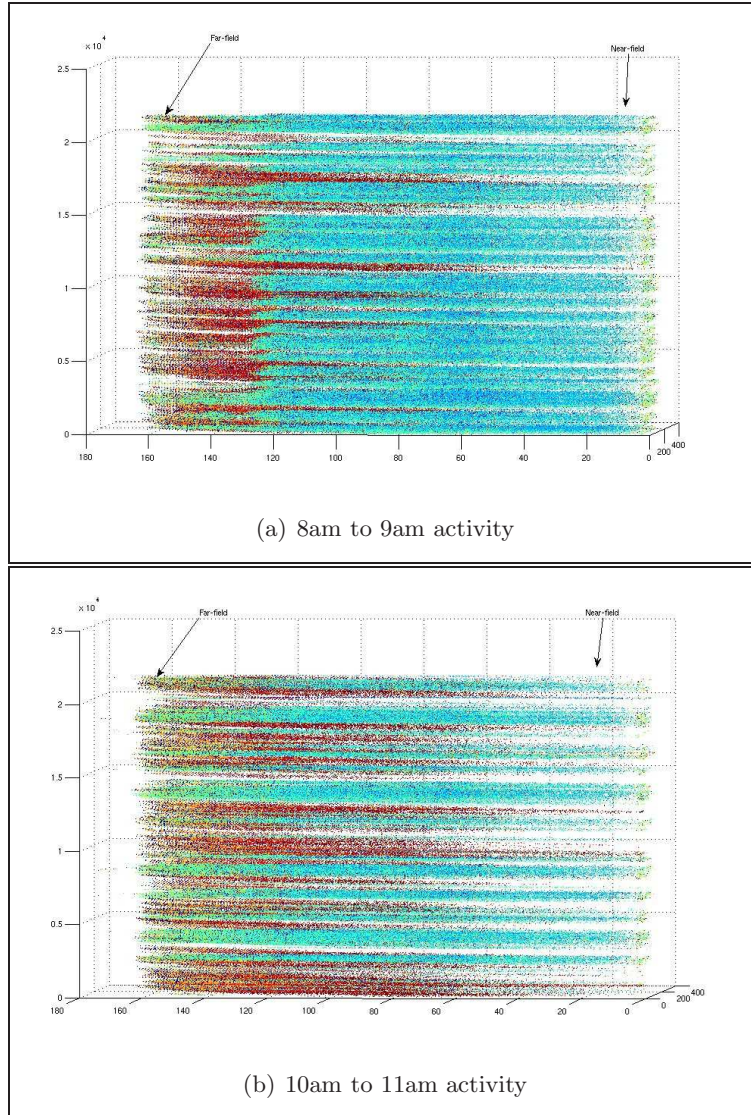


Figure 4-5: 3D scatter plots of activity spread over time across the scene. The plots have been oriented so as to show the length of the passageway. With reference to figure 4-4, the green and blue colour represents the bulk of the people moving from the far-field to the near-field making their way to the outside of the station while the mixture of red and yellow represents people moving from the near-field to the far-field, that is towards the underground train platform.

#### 4.2.2.1 Abnormal video description

The subfigures in Figure 4-6 are some snapshots depicting the kind of abnormalities staged. Figure 4-6(a) shows the outlined actors who stage primarily two abnormal crowd events, explained shortly. The abnormalities staged are few but they aim to violate the general observed flows of motion and the violations caused in the activity descriptors must be detected.

The motion observed in this scene is that of people movement with only few instances of small groups of people (3 or 4) standing still for at most 2 to 3 minutes. The two staged abnormalities aim to violate the normal movement speeds and the normal movement direction.

- **Abnormality 1 (Speed violation) - Large group(s) of people standing still for a couple of minutes:** In this staging, as seen in Figures 4-6(a) and 4-6(b) the large number of actors remain still with little motion for approximately 2 to 3 minutes. This small window of time would collect significant counts for bins showing zero or small pixel displacements. If such behaviour were a normality and previously seen in training data, then it would be classified as normal. This staging ought to create a mismatch between the model  $N^M$  and the test  $\nu^T$  resulting in an abnormality. The direction histograms may cause a mismatch as well.
- **Abnormality 2 (Direction violation) - A large group of people moving about in non-standard directions:** The standard or observed direction of motion is simply back and forth along the passageway coupled with occasional meandering routes that people take. Crowd abnormalities violating direction could take up any form such as bottlenecks, crisscross chases, etc. all of which may not be individually known, but would show deviation from the model histogram by way of higher non-coinciding bin counts. Therefore a simple motion sequence was staged to disrupt the model histogrammic structure in order to verify the modelling scheme. As per Figures 4-6(c) and 4-6(d) the actors walk perpendicular to the transit motion, that is along the breadth of the passageway. Such perpendicular motion is performed in the far-field and closer to the



(a) The people outlined in white are the actors staging abnormal events. Note actors standing still whilst normal motion is underway.



(b) Actors staging abnormal event of standing still in a high transit area.



(c) Actors moving perpendicular the normal flow of motion.



(d) Actors moving perpendicular to the normal flow of motion with activity starting again in the far-field.

Figure 4-6: Images of the staged Abnormal events at a Railway Passageway. Images (a) and (b) show the abnormal event standing still in a high transit area while images (c) and (d) show abnormal motion directions.

near-field. The snapshots show clean images without usual crowd motion running past the actors in order to visualise the motion. However, crowds of people did pass by during this sequence thus the sequence does not simply generate a completely different direction distribution for  $\theta^T$ , but the usual plus new increased bin counts.

## 4.3 Experiments and Ground-truthing

The experiment details are listed below.

### 4.3.1 Experiment 1: Testing path learning

#### 4.3.1.1 Description

Paths are learned using the technique outlined in Section 3.6.1. Two samples from both datasets are used to generate paths and they are checked against manually segmented scenes. Path segmentation depends on two factors which need to be empirically derived. They are (1) the spatial resolution (window size) used for region labelling and (2) The choice of the binning scheme (optimum number of bins in histogram) used for the direction map and the speed map so as to obtain paths closest to the ground truths. The empirical discovery of the optimum spatial resolution and the binning scheme is summarised in section 4.3.1.3. Once these values are known, they are used for path detection.

#### 4.3.1.2 Ground-truthing of paths

The manual segmentations are performed based on observation of changing direction and displacement of optic flow vectors across the scene as seen in Figures 4-7(a) and 4-7(b). Figure 4-7(b) is an example snapshot of the motion vectors. The empty spot in the lower

right corner is void of vectors only in this image, but is well populated for the entirety of the samples used.

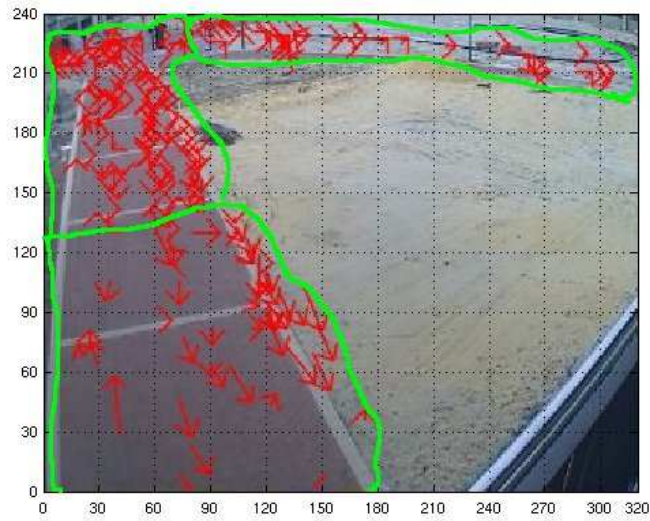
#### 4.3.1.3 Determination of Spatial Resolution and Histogram binning schemes

The task is to determine the spatial resolution that provides the most effective/optimum direction map and speed map which when superimposed upon each other (union operation), provides activity regions closest to the ground truths. The empirical tests carried out are summarised in Figures 4.2 and 4.3 for the direction and speed maps respectively. For each spatial resolution (along rows) extreme binning schemes are displayed (along columns).

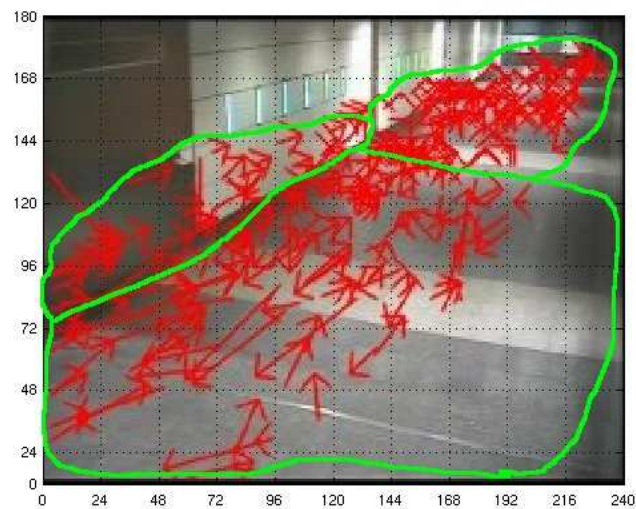
In Figure 4.2 two extreme histogram binning schemes (4 and 16 directions) are shown at different spatial resolutions. Low resolution implies each grid block is sized such that it captures few motion vectors while a higher resolution captures a larger number of motion vectors. It can be seen that a spatial resolution of  $11 \times 8$  and 16 direction bins bearing in mind perspective changes, most effectively determines the direction map.

Figure 4.3 depicts the same principle explained above regarding resolution and extreme binning schemes for displacement/speed. The displacement in pixels occurring in each block of the spatial grids are summarised by 3 and 15 bins for different resolutions. The middle row matches closely with displacement sizes as seen in the ground truthed Figure 4-7(a).

The resolution and binning schemes were found to work well with the railway passageway dataset, but only the university dataset was used to summarise the empirical derivation to avoid information overload (figures). The results of the path detection are based on the conclusive values as seen in Figures 4.2 and 4.3.



(a) University paths



(b) Railway passageway paths

Figure 4-7: Manually created ground truthed paths based on motion similarity. Motion similarity implies each path has motion vectors of a mixture of similar directions and a mixture of similar speeds.



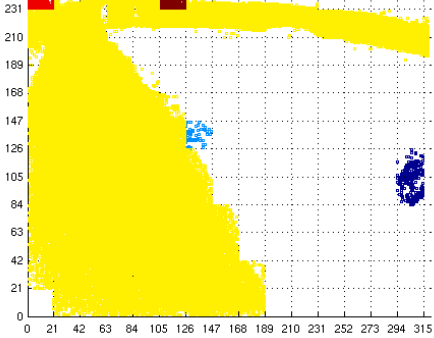
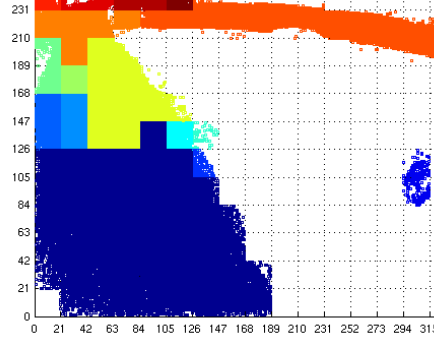


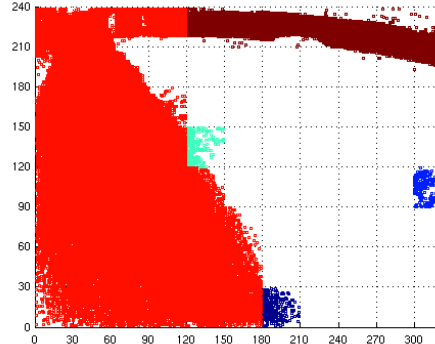

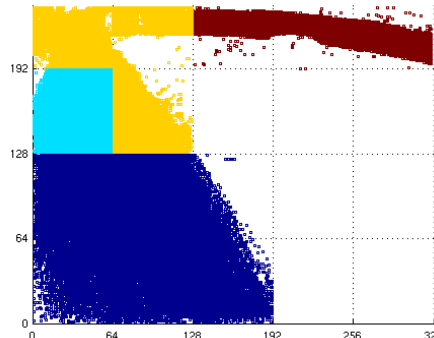
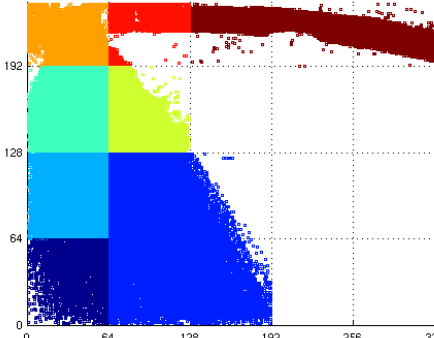


Spatial Resolution (Cols×Rows)	Histogram bins used	
15×12	 <p>A scatter plot showing movement paths on a grid from (0,0) to (315,231). The paths are colored into 4 bins: dark blue, light blue, yellow, and red. The dark blue area is the largest, covering the bottom-left quadrant.</p>	 <p>A scatter plot showing movement paths on a grid from (0,0) to (315,231). The paths are colored into 16 bins, showing a much more granular distribution of directions compared to the 4-bin plot.</p>
Low	4 bins 	16 bins 
11×8	 <p>A scatter plot showing movement paths on a grid from (0,0) to (300,240). The paths are colored into 16 bins. The plot is labeled 'Conclusive - 16 bins'.</p>	
Conclusive	Conclusive - 16 bins 	
5×4	 <p>A scatter plot showing movement paths on a grid from (0,0) to (320,192). The paths are colored into 4 bins: dark blue, light blue, yellow, and red.</p>	 <p>A scatter plot showing movement paths on a grid from (0,0) to (320,192). The paths are colored into 16 bins, showing a more granular distribution of directions.</p>
High	4 bins 	16 bins 

Table 4.2: Summary of empirically obtained or conclusive/optimum number of direction bins vs optimum spatial resolution that accurately captures direction component of paths. On display are pathways from the university dataset.

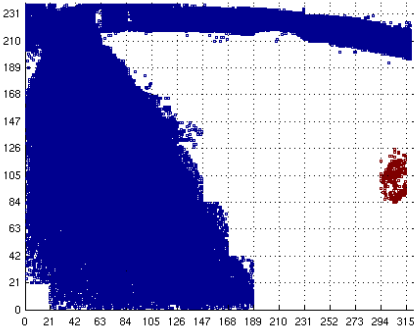
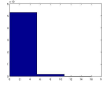
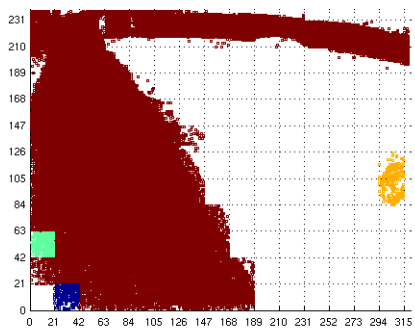
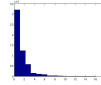
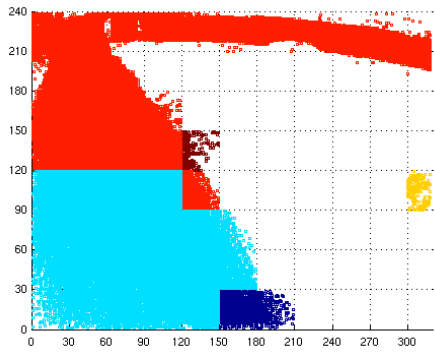
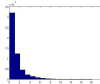
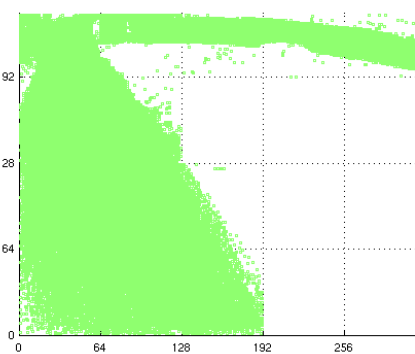
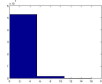
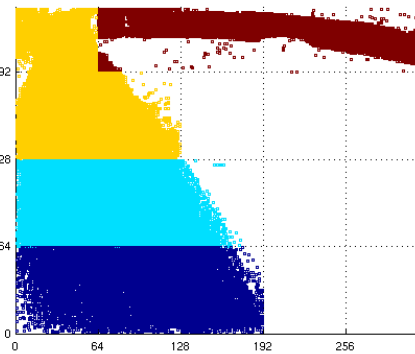
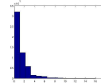
Spatial Resolution (Cols×Rows)	Histogram bins used	
<p>15×12</p> <p>Low</p>	 <p>3 speed bins</p> 	 <p>15 speed bins</p> 
<p>11×8</p> <p>Conclusive</p>	 <p>Conclusive - 16 speed bins</p> 	
<p>5×4</p> <p>High</p>	 <p>3 speed bins</p> 	 <p>15 speed bins</p> 

Table 4.3: Summary of empirically obtained or conclusive/optimum number of speed bins vs optimum spatial resolution that accurately captures speed component of paths. On display are pathways from the university dataset.



#### 4.3.1.4 Samples Used

The following samples listed in Table 4.4 is a subset of the training data and were used for experiment 1.

Sample	Dataset
1 hour approx.	University pathways
Tue 8am-9am 1 hour	Railway passage
Thu 8am-9am 1 hour	Railway passage
Thu 10am-11am 1 hour	Railway passage
All the above 3 hours	Railway passage

Table 4.4: Video samples from training data used for learning pathways. The last sample of 3 hours of the railway passageway constitutes the previous 3 to determine if the path boundaries move while applying the idea of re-enforcing temporal motion over the spatial plane.

#### 4.3.1.5 Parameters Used

The generation of paths requires foreground motion vectors that are extracted from each block of pixels of the image to generate the direction and speed histograms. The following were the parameters used:

##### Background Subtraction

\* By Stauffer and Grimson ([Stauffer and Grimson 2000](#))

\* adaptation rate = 0.01

\* Number of Gaussians = 3

##### Optic Flow

- \* Pyramidal Lucas Kanade ([Bouguet 2000](#))

- \* Number of features: 400

### **Direction Histogram**

- \* range:  $-180^{\circ}$  to  $180^{\circ}$

- \* bins: 16

- \* Region labelling threshold = 0.95. Method of comparison: Normalised cross correlation.

### **Speed Histogram**

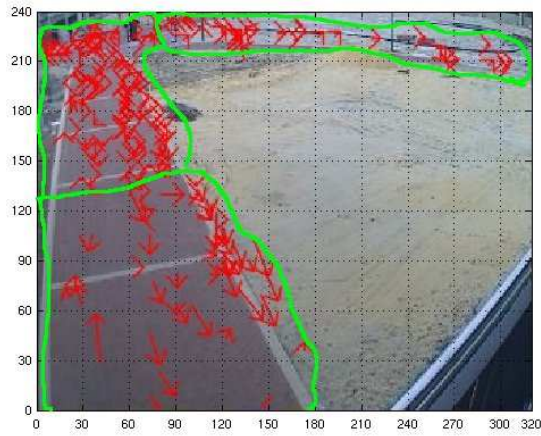
- \* range: 16 bins over the range of the speed values.

- \* Region labelling threshold = 0.95. Method of comparison: Normalised cross correlation.

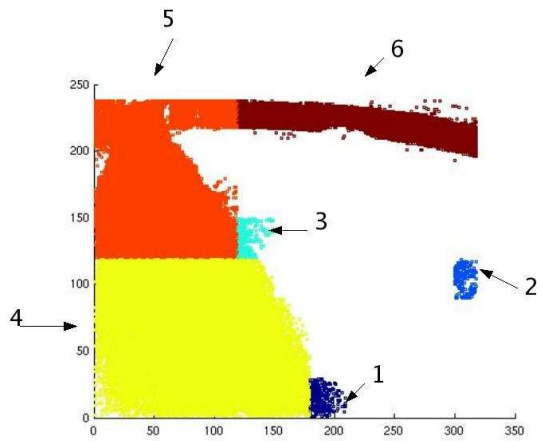
## **4.3.2 Results of path detection: University sample**

Motion similar regions as introduced in Section 3.3 are regions whose unit spaces (along the XY plane) contain columns (along time axis) of similar motion. The ground truthed image is based on motion similarity. Visibly similar motion vectors are grouped together based on changing directions and displacements. The idea of motion similarity does not consider just one direction flow but the collection of motion patterns observed over a unit region over time.

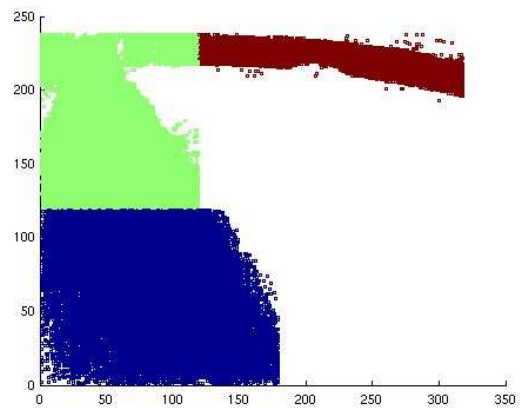
The change in motion across the observed paths is caused by near and far-field effect. As can be seen in the ground truthed Figure 4-8(a), the motion vectors are bordered based on visually perceived change in displacement size and direction. The results in Figure 4-8(b) (containing noise) is an almost exact match. Regions 1 and 3 are regions that are used relatively less and contain just a one-way direction distribution (one way people flow) while regions 4 and 5 contain a uniform distribution of two-way motion. Sufficient training data would inject



(a) University pathways - ground truthed



(b) Detected pathways including noise



(c) Detected pathways after noise thresholding

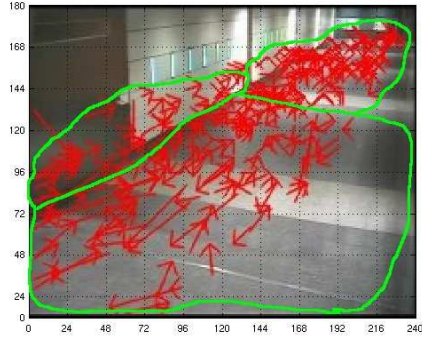
Figure 4-8: Results for experiment 1 (Path detection): University Pathways

two-way motion across regions 1 and 3 after which they ought to merge with regions 4 and 5 respectively.

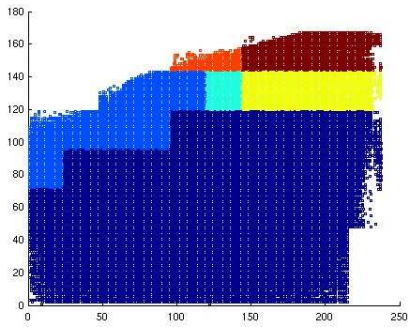
Region 2 contains noise and matches the quantity of usage of regions 1 and 3. Thresholding these results in Figure 4-8(c). The noise thresholding explained in chapter 4 is generically used for both datasets. As the outdoor dataset (University) experiences light changes due to changing time of day and cloud cover, it presents more noise. Therefore additional thresholding is required for outdoor scenes.

### 4.3.3 Results of path detection: Railway Passageway dataset

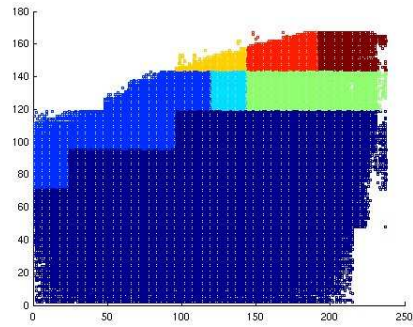
As mentioned above, the ground truthed paths too in Figure 4-9(a) are based on visually similar motion regions. Figures 4-9(b), 4-9(c) and 4-9(d) are paths detected from three samples from the training data while Figure 4-9(e) shows paths obtained by combining the previous three samples. Based on the effect posed by near and far-field phenomenon and the meaning of motion similarity, the one *observed path* (the railway passageway) is divided into approximately 6 regions or paths. As Figure 4-9(e) is representative of the 3 samples, this figure is used to explain the results. With reference to Figure 4-9(e), the scene can be divided into two halves based on the near and far field effect. Regions 1 and 2 in the near-field and regions 3, 4, 5 and 6 in the far-field. Movement along this passageway as seen in the training videos is similar to a two-way street of vehicle traffic where there is no real demarcation but the bulk of the traffic flow demarcates its pathways based on re-enforcement and the ratio it shares with the traffic in the opposite direction. With reference to Figure 4-9(a), the bulk of the motion flows from the far-field to the near-field and therefore occupies a major portion of the width of the passageway, while motion from the near to the far-field moves along leftmost strip of the passageway. Based on this explanation, motion from the far to the near field moves along regions 6, 3, 4 and 1 while motion from the near to the far-field moves along regions 2, 3, 5 and 6. Region 3 and 6 are regions which witness a confluence of two-way flow and hence distinguish themselves from the other regions which have re-enforcing one-way direction distributions.



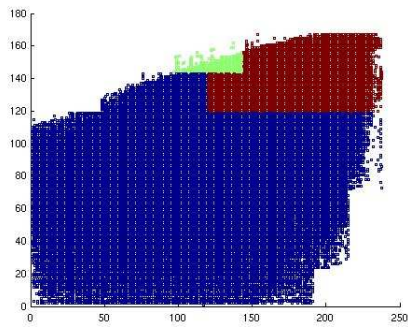
(a) Railway Passageway - ground truthed



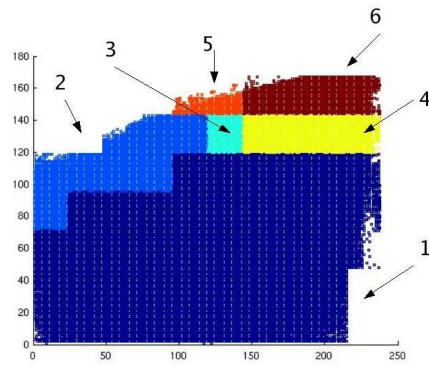
(b) Paths - sample: Tue 8am-9am



(c) Paths - sample: Thu 8am-9am



(d) Paths - sample: Thu 10am-11am



(e) Paths - from combined previous 3 samples

Figure 4-9: Results for experiment 1 (Path Detection): Railway Passageway. As can be seen re-enforcing motion from (b), (c) and (d) provides more detail in (e).

Although the ground-truthed figure, Figure 4-9(a) is segmented based on visual changes, Figure 4-9(e) conforms to this segmentation but revealing greater details. That is, the far-field is split into three regions instead of one. Further, based on the training data, the division of the far-field is somewhat consistent as seen in Figures 4-9(b), 4-9(c) and 4-9(d).

The purpose of this experiment was to establish that the technique devised for determining motion similar regions is consistent and can therefore offer reliability to the system components that depend on it.

#### **4.3.4 Experiment 2: Validation of proposed activity modelling and suitability of temporal resolution**

This section presents the results that demonstrate/establish that the system has the ability to successfully distinguish between similar and dissimilar pairs of activity samples.

##### **4.3.4.1 Description**

This experiment aims at validating the proposed scheme for activity summarisation by comparing similar and dissimilar samples or voxels of activity as described in Section 3.8. If similar or dissimilar samples are in fact found to be so, then this suggests that the proposed scheme decomposes activity sufficiently for comparisons. Successful comparisons also supports/confirms the temporal resolution used. Before presenting these results, as in the previous section, the rationale behind the choice of the temporal resolution used is explained. Temporal resolution is key as a small time window would highlight person details while unaware of crowd trends (unfavourable) while a large time window would suppress person trends and focus on crowd patterns and may even suppress minor abnormalities if they are present. Therefore the optimal resolution is one that is small enough to sufficiently capture crowd trends and emerging abnormalities while large enough to suppress single person trends.

#### 4.3.4.2 Ground-truthing

Two samples are compared by plotting their motion data in a 3D system and colour coding the motion vectors based on motion change. Motion change is simply the process of first binning motion vectors based on direction and then binning each direction bin into different speed bins based on observed speeds. This provides a simplistic non-processed representation of the activity. It is referred to as non-processed as the mass of the data is simply compressed in it's representation by the use of histograms. The data is not scaled or smoothed, etc. Much of the processing in this thesis follows similar steps. This has been done in order to bring out the inherent patterns in the data by a series of simpler representations.

The following are example plots of similar and dissimilar pairs of samples. Ground truthing two samples of activity as similar or dissimilar is an intuitive process and a bird's eye view of the activity is preferred to simply counting people on a frame by frame basis. The plots provide:

- A visual spread of the sample pairs, thus revealing spread of the data.
- The colour coded motion vectors broadly fall into two colours (in this example) explained by two-way motion. Visual inspection of the quantum of masses of each colour in either plot of the pair, shows if these are similar. The quantum of masses may be similar even if they are scattered differently. Figure 4-10 shows two plots of similar activity. The similar colours represents similar binning of motion vectors and visual inspection shows that the quantity of colours in either plot is similar although they are scattered differently (variation). Figure 4-11 shows ground-truthing plots of dissimilar activity.

#### 4.3.5 Determination of Temporal Resolution

The railway passage dataset was used to obtain the appropriate time resolution for comparisons as it contained samples of a full hour (each) which could be easily searched in different

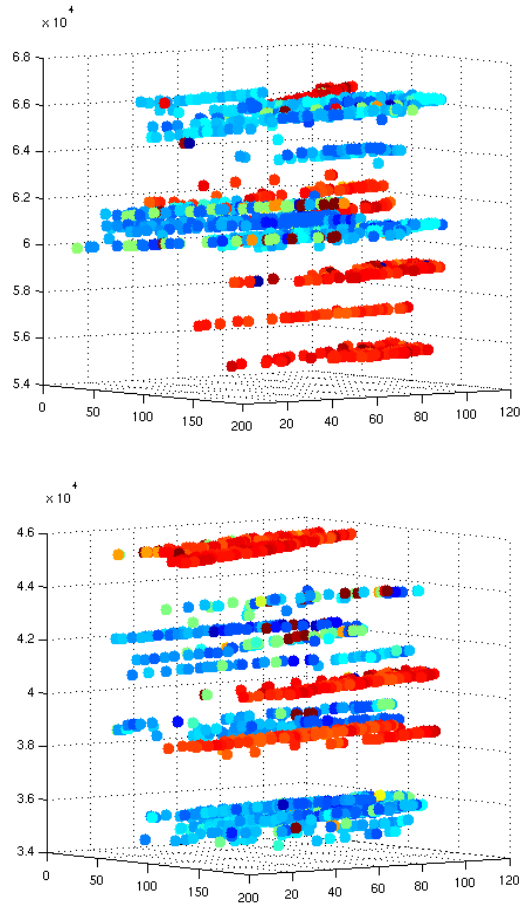


Figure 4-10: Ground truthing example of a **similar pair** of samples taken from the university dataset. It provides a birds eye view of the number of observed directions confirming two-way motion indicated by roughly two colours spread over the XY spatial plane, their quantities in either direction and the way in which they are scattered across equal amounts of time (2 minute samples). This is accurate with what is seen in the video.



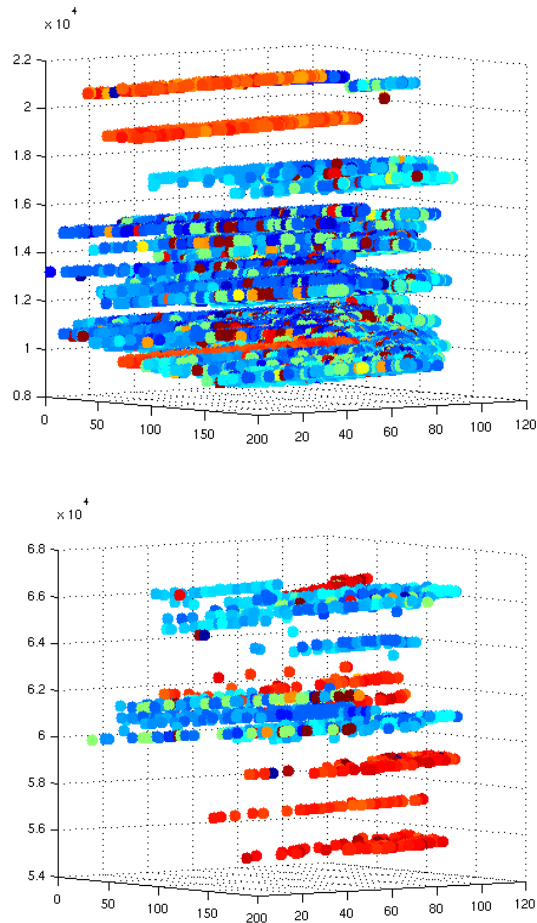


Figure 4-11: Ground truthing example of a **dissimilar pair** of samples taken from the university dataset. It provides a birds eye view of the number of observed directions confirming two-way motion indicated by roughly two colours spread over the XY spatial plane, their quantities in either direction and the way in which they are scattered across equal amounts of time (2 minute samples). This is accurate with what is seen in the video.

resolutions (quantitatively). Tables 4.5, 4.6 and 4.8 show piecewise comparisons between the same pair of similar activity samples (each an hour long) but in different time resolutions. It can be seen that time resolution of 3 minutes (Table 4.5) shows two failed comparisons and naturally so as the time window at that instant could not summarise the generality of the two way activity. However Tables 4.6 (10 min. window), 4.7 (15 min window) and 4.8 (20 min window) capture the general ratio of direction flows over a larger time span. The resolution of 15 minutes is selected as it shows slightly higher accuracy over the 10 minute resolution primarily in the spread results,  $\delta_\rho$  and is of-course smaller than the 20 minute resolution which also shows the hour long samples in a piecewise fashion as similar. Hence, the time resolution of 15 minutes is selected for the experiments 2 and 3.

For experiment two all three activity descriptors of each sample need to be compared as explained in the previous chapter. Each of the descriptors, direction ( $\theta$ ), speed ( $\nu$ ) and spread ( $\rho$ ) is a histogram. The direction and speed histogram use the same binning scheme used in path detection (above) as paths are discovered by collapsing the entire axis of time while direction and speed information for samples are obtained by collapsing a smaller portion of time. The spread histogram is made up of one minute bins. A single minute sufficiently captures the scatter of motion along a path and presents a realistic picture of activity spread over time. The greater the bin-size, the more quickly does the histogram generalise itself such that intuitively differing motion spreads depict as roughly the same histogrammic distribution.

#### 4.3.5.1 Samples Used

The pairs of samples used are listed in Table 4.9.

Temporal Resolution	Time	$\delta_\theta$	$\delta_\nu$	$\delta_\rho$	Status
3 mins	08:00 - 08:03	0.9934	0.9986	0.9351	normal
	08:03 - 08:06	0.9877	0.9847	0.7959	normal
	08:06 - 08:09	0.9758	0.9981	0.8518	normal
	08:09 - 08:12	0.9772	0.9737	0.8320	normal
	08:12 - 08:15	0.9961	0.9974	0.7484	normal
	08:15 - 08:18	0.9920	0.9952	0.9256	normal
	08:18 - 08:21	0.9927	0.9883	0.9347	normal
	08:21 - 08:24	0.9694	0.9944	0.8806	normal
	08:24 - 08:27	0.9928	0.9946	0.8795	normal
	08:27 - 08:30	0.9945	0.9944	0.8802	normal
	08:30 - 08:33	0.9508	0.9990	0.9955	normal
	08:33 - 08:36	0.9866	0.9732	0.8851	normal
	08:36 - 08:39	0.8863	0.9994	0.9592	abnormal
	08:39 - 08:42	0.9725	0.9943	0.9523	normal
	08:42 - 08:45	0.9970	0.9936	0.9695	normal
	08:45 - 08:48	0.9950	0.9994	0.7953	normal
	08:48 - 08:51	0.9949	0.9928	0.9332	normal
	08:51 - 08:54	0.9969	0.9957	0.8294	normal
	08:54 - 08:57	0.4933	0.9860	0.8869	abnormal
	08:57 - 09:00	0.9586	0.9961	0.9668	normal

Table 4.5: Summary comparisons using a resolution of 3 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar

Temporal Resolution	Time	$\delta_\theta$	$\delta_\nu$	$\delta_\rho$	Status
10 mins	08:00 - 08:10	0.9985	0.9985	0.8308	normal
	08:10 - 08:20	0.9993	0.9982	0.8224	normal
	08:20 - 08:30	0.9892	0.9957	0.8501	normal
	08:30 - 08:40	0.9973	0.9960	0.7636	normal
	08:40 - 08:50	0.9911	0.9992	0.7909	normal
	08:50 - 09:00	0.9673	0.9981	0.7948	normal

Table 4.6: Summary comparisons using a resolution of 10 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar

Temporal Resolution	Time	$\delta_\theta$	$\delta_\nu$	$\delta_\rho$	Status
15 mins	08:00 - 08:15	0.9995	0.9979	0.8437	normal
	08:15 - 08:30	0.9974	0.9952	0.8460	normal
	08:30 - 08:45	0.9972	0.9974	0.7470	normal
	08:45 - 09:00	0.9843	0.9994	0.8327	normal

Table 4.7: Summary comparisons using a resolution of 15 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar

Temporal Resolution	Time	$\delta_\theta$	$\delta_\nu$	$\delta_\rho$	Status
20 mins	08:00 - 08:20	0.9994	0.9993	0.8362	normal
	08:20 - 08:40	0.9976	0.9990	0.8337	normal
	08:40 - 09:00	0.9961	0.9998	0.7906	normal

Table 4.8: Summary comparisons using a resolution of 20 mins between Tue 8am-9am and Thu 8am-9am which are ground truthed as similar

Sr No	Category	Dataset	sample 1 time	sample 2 time
1	similar pair 1	Railway passage, path 1	Tue 8am-8:15am	Thu 8am-8:15am
2	similar pair 2	Railway passage, path 1	Tue 8:45am-9am	Thu 8:45am-9am
3	dissimilar pair 1	University, path 4	10am-10:15am	10:15am-10:30am
4	dissimilar pair 2	University, path 5	10:15am-10:30am	10:45am-11am

Table 4.9: Samples used for experiment 2

### 4.3.6 Parameters and settings used

The thresholds set for comparisons for each of the activity descriptors  $\theta$ ,  $\nu$  and  $\rho$  are 0.95, 0.95 and 0.8 respectively as listed in Table 4.11. The threshold for cross-correlation similarity is relaxed for the spread histograms as a number of variations must be considered and appropriately matched as only the alarmingly different ones are of importance. Table 4.10 provides the settings and binning values for the motion descriptor histograms.

Histogram Properties	Direction Histogram $\theta$	Speed Histogram $\nu$	Spread Histogram $\rho$
Range	$-180^\circ$ to $180^\circ$	16 bins over the observed speeds	Width represents duration of sample
Number of Bins	16	16	15 bins @ 1minute of data/bin
Bin Size	$22.5^\circ$	–	1 minute
Threshold Used	$\tau_\theta = 0.95$	$\tau_\nu = 0.95$	$\tau_\rho = 0.8$

Table 4.10: Binning scheme used for storing motion information and cross correlation thresholds used for comparisons in experiments 2 and 3.

Sample Pairs	$\delta_\theta$	$\delta_\nu$	$\delta_\rho$
Similar pair 1	0.9994	0.9974	0.8091
Similar pair 2	0.9899	0.9995	0.8181
Dissimilar pair 1	<b>0.9066</b>	0.9935	<b>0.6850</b>
Dissimilar pair 2	<b>0.8405</b>	0.9716	<b>0.1411</b>
Thresholds: $\tau_\theta = 0.95$ , $\tau_\nu = 0.95$ and $\tau_\rho = 0.8$			

Table 4.11: Normalised cross-correlated values of the sample comparisons,  $\delta_\theta$ ,  $\delta_\nu$  and  $\delta_\rho$

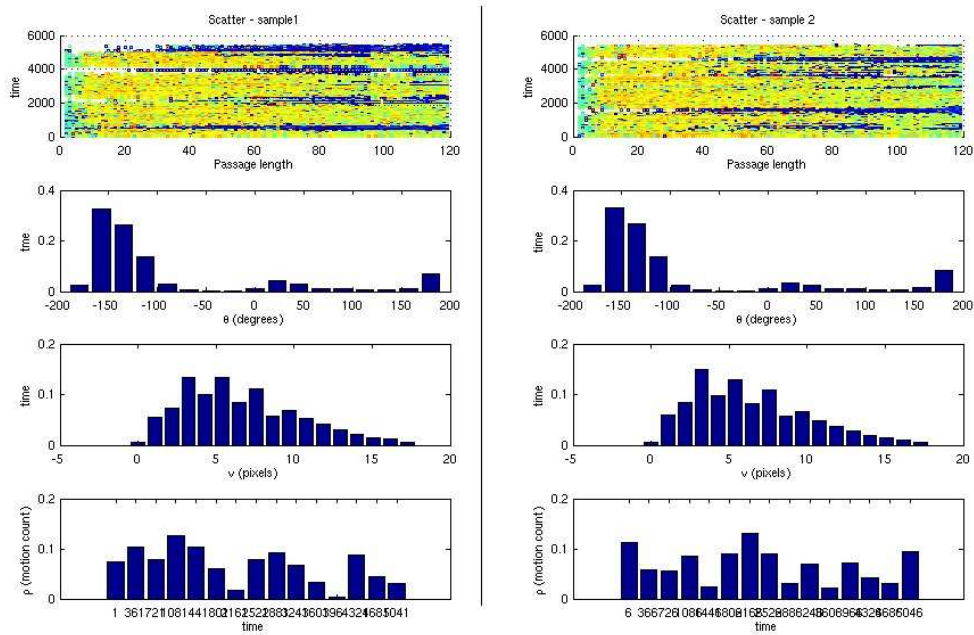


Figure 4-12: Comparison decomposition of similar samples - 1. First row: 3D Scatter plots oriented to show the length of the railway passage way. Row 2: direction histograms. Row3: Speed histograms. Row 4: Spread histogram. Each bin in Row 4 histogram represents unit time and the number of bins completes the time sample. **The X axis of Row 4 (unreadable) is composed of 15 1 minute bins.**

### 4.3.7 Results: Comparison of similar samples

Figure 4-12 is a decomposition of the similar pairs of samples. It is broadly divided into 2 columns, one for each sample. The first sub-figure in a column is the 3D scatter plot providing a visual feel for the data, spread over time and mixture of directions present within. Thereafter are the histograms for each of the activity descriptors, direction, speed and spread. As can be seen these are visually similar and their **normalised** cross correlated values are listed in Table 4.11.

The two sample are of the same time duration across Tuesday and Thursday and as can be seen are almost identical in  $\theta$  and  $\nu$  with slight difference in the spread histogram as seen in the 12<sup>th</sup> minute .

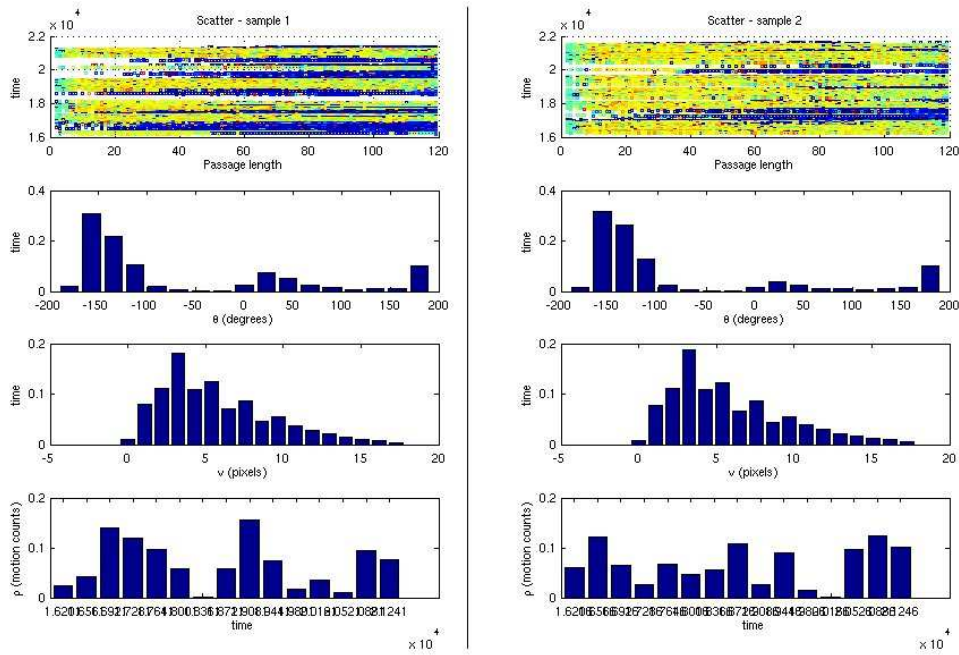


Figure 4-13: Comparison decomposition of similar samples - 2. First row: 3D Scatter plots oriented to show the length of the railway passage way. Row 2: direction histograms. Row3: Speed histograms. Row 4: Spread histogram. Each bin in Row 4 histogram represents unit time and the number of bins completes the time sample. **The X axis of Row 4 (unreadable) is composed of 15 1 minute bins.**

Figure 4-13 describes the comparison between samples again of the same time-of-day across different days (8:45am-9am Tue & Thu). The sub-figure plots show that the left plot contains more blue motion than the right and this appears so due to the view orientation. However, their direction histograms show similar quantities for motion distribution and is invariant to the way the flows are scattered across the respective samples. The speed histograms appear identical with a cross-correlation of 0.9995 (refer Table 4.11) while the spread histograms are very similar with a cross-correlated similarity of 0.8181.

### 4.3.8 Results: Comparison of dissimilar samples

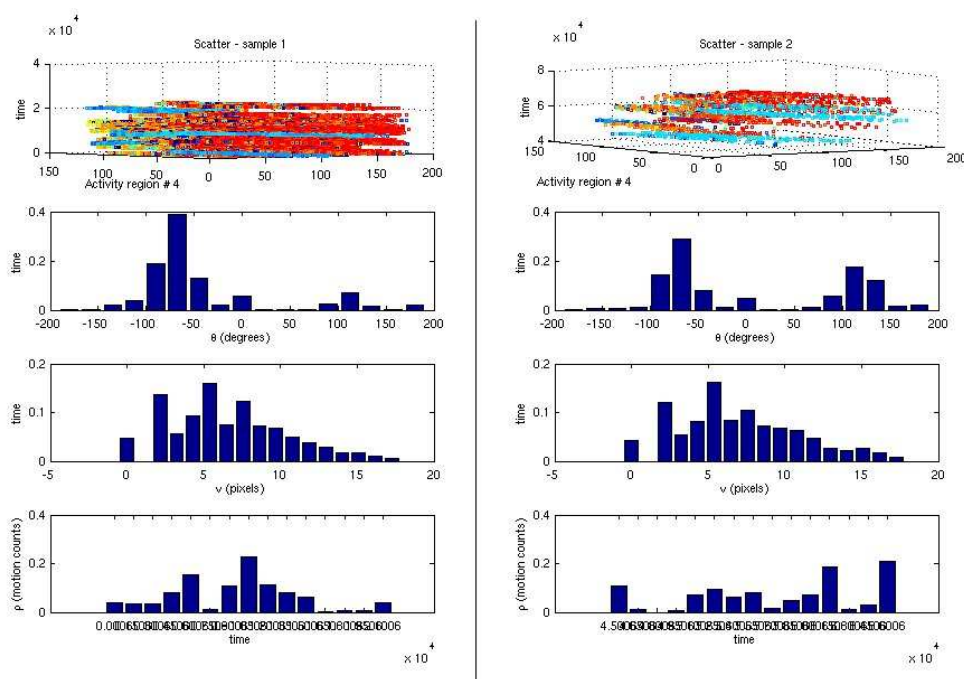


Figure 4-14: Comparison decomposition of dissimilar samples - 1. First row: 3D Scatter plots of activity. Row 2: direction histograms. Row 3: Speed histograms. Row 4: Spread histogram. Each bin in Row 4 histogram represents unit time and the number of bins completes the time sample. **The X axis of Row 4 (unreadable) is composed of 15 1 minute bins.**

With reference to Figure 4-14, the dissimilar pair of activity, the scatter plots present this dissimilarity intuitively. The speed histograms are similar but the direction histograms are not. The  $\theta$  histogram in the left column shows a dominant cluster (bins) of directions implying



mostly one-way motion while the right column counterpart shows two similarly shaped well spaced distributions implying two-way motion. This compares as a cross correlation of 0.9066 which is still lesser than the threshold of 0.95. The two direction histograms are similar in shape with respect to the first half of the histogram from bins  $-180^\circ$  to  $0^\circ$ , hence the correlation is high. The spread is somewhat dissimilar such that  $\rho$  in the right column lacks the central bell shape as seen in the left one. Hence, a cross-correlation of 0.6850.

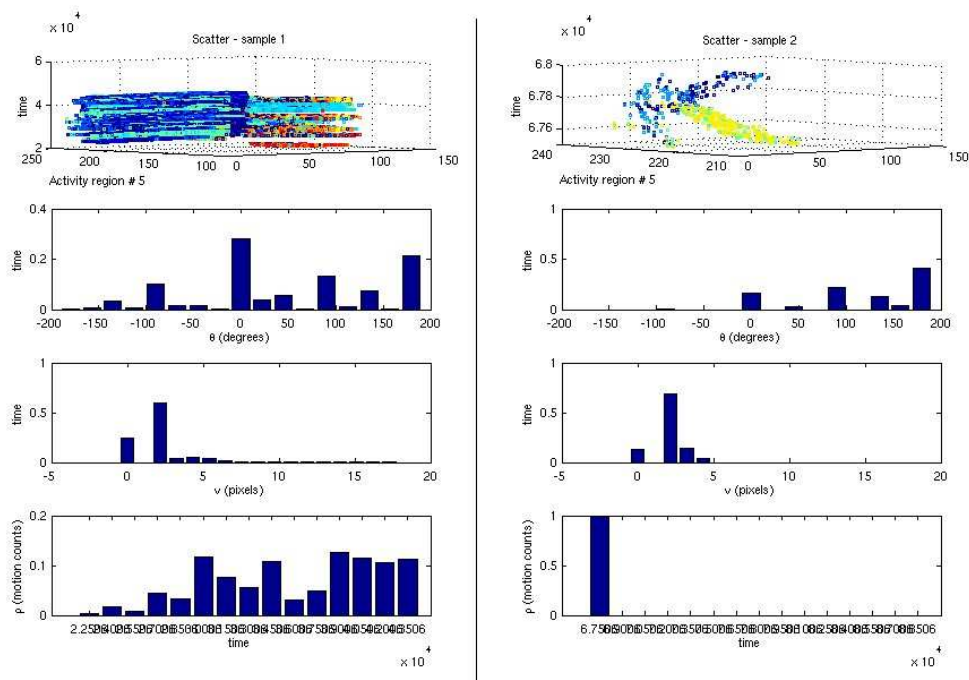


Figure 4-15: Comparison decomposition of dissimilar samples - 2. First row: 3D Scatter plots of activity. Row 2: direction histograms. Row3: Speed histograms. Row 4: Spread histogram. Each bin in Row 4 histogram represents unit time and the number of bins completes the time sample. **The X axis of Row 4 (unreadable) is composed of 15 1 minute bins.**

Figure 4-15 is an example of 2 samples that are significantly different in motion masses yet present similar speeds, close to similar direction distributions but obviously because of differing quantities, the spread histogram is simply dissimilar with a cross-correlation of 0.1411.

In this experiment, similar samples have found to be similar and dissimilar samples have been found to be dissimilar. The purpose of this experiment was to establish that the proposed scheme for modelling activity is parametrised ( $\theta$ ,  $\nu$  and  $\rho$ ) sufficiently such that it is invariant

to the variations presented by scattered motion flows and can capture the intuitive similarity or dissimilarity presented by the samples.

### 4.3.9 Experiment 3: System performance by way of test sample classification with the trained model

#### 4.3.9.1 Description

To test the trained model for its classification outcome for known normal and abnormal test samples. The samples listed in Table 4.12 were used. Reasonable success in classifying the known normal and abnormal samples suggests that the model is derived from sufficient variations of activity and that the system as a whole works.

#### 4.3.9.2 Ground-truthing

The same scheme as used in experiment 2 applies for experiment 3. Motion data for the (p,t) used are plotted individually and then compared against the plot of the test data in addition to visual inspection of the actual videos.

#### 4.3.9.3 Samples used

Day	Duration		Total	Type
Tue	1 × 8am-9am	1 × 10am-11am	2 hours	train
Thu	1 × 8am-9am	1 × 10am-11am	2 hours	train
Wed	1 × 8am-9am	1 × 10am-11am	2 hours	test normal
Wed	–	1 × 10am-10:15am	15 mins	test abnormal

Table 4.12: Samples used in experiment 3 (Railway passage only)

### 4.3.10 Parameters and settings used

The thresholds set for comparisons for each of the activity descriptors  $\theta$ ,  $\nu$  and  $\rho$  are 0.95, 0.95 and 0.8 respectively as listed in Table 4.11. The threshold for cross-correlation similarity is relaxed for the spread histograms as a number of variations must be considered and appropriately matched as only the alarmingly different ones are of importance. Refer Table 4.10 for the settings used.

#### 4.3.10.1 Results of experiment 3

The results of the normal and abnormal test data are presented in Tables 4.13 and 4.14 respectively. Each value in the tables represents how similar the summarised motion descriptor (direction or speed or spread histogram) for a given path (row) for a given period of time (column) is to the model histograms for the same path during the same time of day. That is, each entry is a cross correlated value for a given (p,t) of the sample with the corresponding (p,t) of the model  $\Lambda$ . Refer Figure 4-9(e) for following discussion of paths.

In Table 4.13(normal samples), paths 1 and 2 in the near-field show motion consistency with the quantum and directions of flow via ground truths and video perusal. Motion across path 4 is somewhat spread over the far and near-field and shows consistent results too. But motion along paths 3, 5 and 6 show minor dissimilarities. The lack of similarity is attributed to two factors which are: (1) Motion fields appearing in the far-field are relatively less well defined than those appearing in near-field paths explained by changes in displacement and direction due to the absence of normalisation of scene motion.(2) This could be addressed with sufficient training data. An increase in the training data is likely to generate more variation which may find the legitimately normal motion as normal. There is only one deviation in  $\delta_\nu$  (speed) in path 3 at 10:45am to 11:00am. This is minor and would be resolved with additional training samples. A couple of dissimilarities are observed for  $\delta_\rho$  in paths 3 and 6.

With respect to Table 4.14, as per the video, two abnormalities were staged (refer Section

4.2.2.1). To summarise, they were (a) actors walking back and forth against the direction of normal flow **whilst normal motion flows were underway.** and (b) a large group of actors standing still for an extended period of time during normal motion flows. It was possible to capture stillness in the foreground as the adaptation rate of the background subtraction was set to 0.01. Path 1 contains both abnormalities while paths 2, 3, 4 and 5 contain the abnormality #(a) in which the actors walk against the normal flows of motion.

It can be seen that path 1 shows dissimilarity with respect to the model in all three descriptors.  $\delta_\theta$ , the direction similarity has value of 0.9058 which is still less than the threshold 0.95 given that normal motion was underway during the abnormality. The dissimilarity ought to be greater if abnormality were staged in an empty scene. The speed descriptor similarity,  $\delta_\nu$  shows dissimilarity with a somewhat low value of 0.8910. Again, the actors stood still in path 1 while people were passing by. The dissimilarity could be attributed to the bins representing reduced or zero displacement counts which grew over the extended period of time and gained clear definition. This has not been previously seen in the training data and hence the two signals (one-dimensional histograms) did not correlate well. The density similarity given by  $\delta_\rho$  shows dissimilarity explained by the packed actors in path 1 effecting increased presence of many persons which has not been previously seen in the training data **for that space and time.** The abnormalities detected were due to new histogrammic definitions of direction, speed and density in their respective histograms that have not been previously seen.

Motion over paths 2 and 3 referred to the actors walking back and forth against the flow of directions. They moved at similar speeds as the commuters and this is reflected as similar in the cross correlated value of 0.9650 for path 2. The table shows dissimilarity for the direction and spread descriptor while similarity for the speed. Path 3 as per Figure 4-9(e) lies in the region which witnesses both direction flows of motion and the increase in density is attributed to a clump of persisting motion, the actors who were walking back and forth.

Paths 3, 4 and 5 witnessed the abnormal motion (perpendicular motion flow) and this was detected by way of dissimilarity in the direction descriptor for all three paths. All three paths have normal speeds as the actors moved at normal speeds. Path motion for paths 3 and 5

show spread dissimilarity while path 4 does not. This is in agreement with the video context that people use paths 6, 4 and 1 constituting one way motion from the far to the near field while few people use paths 2, 3, 5 and 6 constituting one way motion from the near to the far field. Hence path 4 is used to observing high motion counts and therefore the spread was similar to the model. No abnormalities were staged in path 6 and the irregular density  $\delta_\rho$  at 0.7436 is attributed to the lack of training data.

Paths		Cross Correlation values of 15 min samples							
		8am		9am		10am		11am	
Path 1	$\delta_\theta$	0.9974	0.9987	0.9980	0.9987	0.9816	0.9750	0.9980	0.9945
	$\delta_\nu$	0.9831	0.9897	0.9936	0.9933	0.9961	0.9992	0.9988	0.9857
	$\delta_\rho$	0.7545	0.8819	0.8916	0.9104	0.7042	0.8556	0.9129	0.9365
Path 2	$\delta_\theta$	0.9911	0.9990	0.9897	0.9932	0.9739	0.9574	0.9982	0.9890
	$\delta_\nu$	0.9863	0.9959	0.9933	0.9920	0.9862	0.9956	0.9983	0.9949
	$\delta_\rho$	0.8975	0.9511	0.7901	0.8540	0.8224	0.8384	0.8276	0.8352
Path 3	$\delta_\theta$	0.9906	0.9984	0.9917	0.9841	0.9703	0.9338	0.9931	0.8292
	$\delta_\nu$	0.9965	0.9996	0.9934	0.9970	0.9875	0.9956	0.9992	0.9162
	$\delta_\rho$	0.8715	0.9334	0.8261	0.8744	0.8580	0.7905	0.7794	0.6535
Path 4	$\delta_\theta$	0.9960	0.9988	0.9942	0.9921	0.9643	0.9599	0.9962	0.9946
	$\delta_\nu$	0.9967	0.9998	0.9986	0.9926	0.9977	0.9987	0.9998	0.9965
	$\delta_\rho$	0.8014	0.8972	0.8634	0.9060	0.7611	0.9127	0.8518	0.9326
Path 5	$\delta_\theta$	0.9979	0.9988	0.9904	0.9982	0.9690	0.9769	0.9914	0.8919
	$\delta_\nu$	0.9992	0.9995	0.9953	0.9977	0.9945	0.9980	0.9994	0.9844
	$\delta_\rho$	0.7985	0.8554	0.7311	0.6945	0.8439	0.6339	0.7829	0.7011
Path 6	$\delta_\theta$	0.9937	0.9991	0.9676	0.9049	0.9581	0.9393	0.9909	0.9965
	$\delta_\nu$	0.9999	0.9998	0.9980	0.9970	0.9965	0.9984	1.0000	0.9999
	$\delta_\rho$	0.8634	0.8403	0.7262	0.8532	0.8064	0.8355	0.8970	0.9594

Table 4.13: Results of normal test samples with the trained model

Paths	Cross Correlation values of 15 min samples	
	10:00am - 10:15am	
Path 1	$\delta_\theta$	0.9058
	$\delta_\nu$	0.8910
	$\delta_\rho$	0.7336
Path 2	$\delta_\theta$	0.7359
	$\delta_\nu$	0.9650
	$\delta_\rho$	0.7283
Path 3	$\delta_\theta$	0.7137
	$\delta_\nu$	0.9937
	$\delta_\rho$	0.7209
Path 4	$\delta_\theta$	0.9399
	$\delta_\nu$	0.9920
	$\delta_\rho$	0.8389
Path 5	$\delta_\theta$	0.8278
	$\delta_\nu$	0.9986
	$\delta_\rho$	0.5361
Path 6	$\delta_\theta$	0.9753
	$\delta_\nu$	0.9997
	$\delta_\rho$	0.7436

Table 4.14: Results of abnormal test samples with the trained model

## 4.4 A Discussion of Weaknesses of the System

At the outset, it has been stated that this system is targeted at crowd anomalies and not single person ones. However, based on the proposed 3D modelling scheme and choice of pre-processing techniques used, the following scenarios pose as the weaknesses of the system. The abnormalities spoken of herein were described earlier in the chapter. Some of the scenarios were not tested for but have a sound theoretical justification.

### 4.4.1 Single person anomalies

A single person vandalising a wall translates as the presence of to-and-fro motion vectors along a wall which is similar to people passing by that wall. If this behaviour is uncommon given a time of day, then such activity is likely to be detected if it persists for 4 to 5 minutes. However, a shorter duration of 1 minute would go undetected as it could be compared to occasional passer-bys.

### 4.4.2 A group of people standing still

A group of people standing still would cause the flow of motion to curve around them thus changing the geometry of the flow. The proposed scheme is invariant to changes in flow geometry but would detect a sizeable spatial region with temporally increasing counts of zero length motion vectors. If this motion was commonly present in the training data, then such behaviour would be deemed normal. Such a scenario was enacted by our actors in the railway passage dataset in the near-field of the camera view for approximately 3 minutes and was successfully detected. However, this may not work in every situation as the adapting background model would be seeking to label motionless or unchanging pixel colours as the background. If people were to stand still for long periods of time, the change in flow geometry would not be detected, but a loss in foreground would occur which results in a reduction in

quantity of foreground motion. Once again, if this is a significant reduction in quantity with respect to the training data, then the abnormality would be detected.

#### **4.4.3 Anomalies across discrete time windows**

In a situation where a small length anomaly is spread across time windows, it likely that it would have gained little or no definition via the re-enforcing motion and hence would not be detected.

#### **4.4.4 Minimum time needed to detect an anomaly**

Given the proposed system of re-enforcing motion patterns, a scenario requires time to gain definition for a meaningful comparison. Most of the enactments in the railway passage dataset were three minutes long with a medium number of people/actors (approximately 15 people) in groups of 3 or 4. The time needed to detect an anomaly depends upon the number of participants, the extent of opposing motion and the duration. One minute long abnormalities went undetected.

### **4.5 Summary**

This chapter described the experiments used to validate the system for modelling crowd activity. The datasets used and the methods of ground truthing were also explained. Experiments 1 and 2 have successfully shown that the system components, namely path detection and temporal comparisons work consistently and that the scheme used to decompose activity so as to conduct comparisons works. Experiment 3 shows that the trained model with just a small amount of training data classifies activity appropriately with few false positives and that additional training data ought to resolve this problem. The activity in a single sam-



ple of 15 mins is treated as six separate anomalies as the abnormal staging spreads across the six different paths. In each case the abnormality was identified adequately with no false negatives. The next chapter presents the conclusion and suggestions to improve the current system.

## Chapter 5

# Conclusion

The goal of this thesis was to devise an approach to learn normal spatio-temporal patterns in crowd motion so as to detect the abnormal. Each of the sub aims as listed in Section 1.3 have been solved as shown in Chapter 3, the methodology and validated as shown in the previous results' chapter.

Part of the motivation behind this work was to make it scalable to most public locations. The factors that prevent this have been observed to be non-preferred camera placements, views of a combination of near and far-fields causing blobs to change size as they propagate through the scene. Existing work use simplistic datasets to which single/global thresholds may be applied which would not work well with complex scene structure. The approach used in this thesis, namely the way in which an *observed path* is segmented into sub paths such that the resultant path contains spatial and temporal similar data seeks to facilitate the scalability of the approach. Moreover, inspite of localised definitions, there are no user defined thresholds for model comparison. The motion descriptors are stored as one-dimensional histograms (signals) and are compared with the resulting test signal through the well established technique of cross-correlation with thresholds that show high confidence indicating similarity.

## 5.1 Future Work

A few improvements to this thesis would be advantageous in the future for greater scalability. They are as follows:

- The areas to improve in this thesis would be to learn temporal boundaries of activity instead of setting discrete 15 minute chunks of time. This would make the system more generic for learning a location's individual temporal activity patterns. This could be achieved by drawing ideas from [Uncu et al. \(2006\)](#) and [Zheng et al. \(2010\)](#) who have similar grid-based approaches to clustering spatial data with differing densities.
- Automated scene calibration or achieving normalisation of scene motion would help in ensuring that the segmented paths are indeed the same as the observed paths. If a path lay both in the near and far field of a scene such that the proposed algorithm would segment it into a number of small paths, then any abnormalities occurring within the smaller paths may go unnoticed as the spatial resolution would be reduced. Scene normalisation would enable learning spatial patterns across the length of the whole observed path making the conclusions of the comparisons more reliable.
- Finally, it would be desirable to not assume paths once learned to be the absolute paths. The system should be able to deal with changing motion patterns implying the way people use a region may change within the day and this temporal change of path definitions ought to be reflected in the model. For example, a playground may be used for parades in the morning, random people motion in the afternoon and well defined 3 or 4 way motion flows in the evenings. It would be ideal to maintain a number of path models based on the time of day to make the system dynamic or consistently learning.

# Bibliography

- Andrade, E. L., S. Blunsden, and R. B. Fisher (2006). Modelling crowd scenes for event detection. In *ICPR 2006. 18th International Conference on Pattern Recognition, 2006*, Volume 1, Hong Kong, pp. 175–178. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed March, 2007).
- Avidan, S. (2001). Support vector tracking. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, Volume 1, pp. I-184–I-191 vol.1.
- Bar-Shalom, Y. and T. Foreman (1988). *Tracking and Data Association*. Academic Press Inc.
- Bertalmio, M., G. Sapiro, and G. Randall (2000, Jul). Morphing active contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(7), 733–737.
- Black, M. J. and P. Anandan (1993, May). A framework for the robust estimation of optical flow. In *Proceedings of the Fourth International Conference on Computer Vision, 1993.*, pp. 231–236. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed June, 2007).
- Black, M. J. and A. D. Jepson (1998). Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *International Journal of Computer Vision*, Volume 26, pp. 63–84.
- Boghossian, B. A. and S. A. Velastin (1998). Real-time motion detection of crowds in video signals. In *IEE Colloquium on High Performance Architectures for Real-Time Image Processing*, London, UK, pp. 1–6. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed January, 2007).
- Boghossian, B. A. and S. A. Velastin (1999). Motion-based machine vision techniques for the

- management of large crowds. In *Proceedings of ICECS '99 - The 6th IEEE International Conference on Electronics, Circuits and Systems, 1999*, Volume 2, Pafos, pp. 961–964. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed February, 2007).
- Bouguet, J.-Y. (2000). Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. Intel Corporation, Microprocessor Research Labs. [http://robots.stanford.edu/cs223b04/algo\\_tracking.pdf](http://robots.stanford.edu/cs223b04/algo_tracking.pdf) (accessed August, 2007).
- Broida, T. J. and R. Chellappa (1986, Jan.). Estimation of object motion parameters from noisy images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8*(1), 90–99. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed August, 2008).
- Brostow, G. J. and R. Cipolla (2006). Unsupervised bayesian detection of independent motion in crowds. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, Volume 1, New York, US, pp. 594–601. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed March, 2007).
- Canny, J. (1986, Nov.). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8*(6), 679–698. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed Jan 2009).
- Casas, J. R., A. P. Sitjes, and P. P. Folch (2005). Mutual feedback scheme for face detection and tracking aimed at density estimation in demonstrations. In *IEE Proceedings on Vision, Image and Signal Processing*, Volume 152, pp. 334–346. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed March, 2007).
- Comaniciu, D., V. Ramesh, and P. Meer (2003, May). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*(5), 564–577.
- Davies, A. C., J. H. Yin, and S. A. Velastin (1995, Feb). Crowd monitoring using image processing. *Electronics & Communication Engineering Journal 7*(1), 37–47. IEEEExplore database. <http://ieeexplore.ieee.org> (accessed April, 2007).
- Efros, A. A., A. C. Berg, G. Mori, and J. Malik (2003). Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision ICCV '03*, Washington, DC, USA, pp. 726. IEEE Computer Society. ACM Digital Library. <http://portal.acm.org> (accessed June, 2009).

- Heikkila, M., M. Pietikainen, and J. Heikkila (2004). A texture based method for detecting moving objects. In *British Machine Vision Conference*, Kingston University, London, pp. 187–196.
- Horn, B. K. and B. G. Schunck (1980). Determining optical flow. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA. ACM Digital Library. <http://portal.acm.org> (accessed April, 2008).
- Huttenlocher, D. P., J. J. Noh, and W. J. Rucklidge (1993, May). Tracking non-rigid objects in complex scenes. In *Proceedings of the Fourth International Conference on Computer Vision, 1993.*, pp. 93–101.
- Irani, M. and P. Anandan (1999). About direct methods. In *Workshop on Vision Algorithms*, pp. 267–277. Springer Berlin / Heidelberg. <http://link.springer.de/link/service/series/0558/bibs/1883/18830267.htm> (accessed July, 2008).
- Isard, M. and A. Blake (1998, August). Condensationconditional density propagation for visual tracking. *International Journal of Computer Vision* 29(1), 5–28.
- Kang, J., I. Cohen, and G. Medioni (2004, Aug.). Object reacquisition using invariant appearance model. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Volume 4, pp. 759–762.
- Khan, S., O. Javed, Z. Rasheed, and M. Shah (2001). Human tracking in multiple cameras. In *Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001.*, Volume 1, pp. 331–336 vol.1. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed January, 2008).
- Lucas, B. D. and T. Kanade (1981, April). An iterative image registration technique with an application to stereo vision (DARPA). In *Proceedings of the 1981 DARPA Image Understanding Workshop*, pp. 121–130. [http://www.ri.cmu.edu/pub\\_files/pub3/lucas\\_bruce\\_d\\_1981\\_2/lucas\\_bruce\\_d\\_1981\\_2.pdf](http://www.ri.cmu.edu/pub_files/pub3/lucas_bruce_d_1981_2/lucas_bruce_d_1981_2.pdf) (accessed July, 2008).
- Marana, A. N., S. A. Velastin, L. F. Costa, and R. A. Lotufo (1997). Estimation of crowd density using image processing. In *IEE Colloquium on Image Processing*

- for *Security Applications*, Volume 1, London, UK, pp. 1–8. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed February, 2007).
- McIvor, A. M. (2000). Background subtraction techniques. In *In Proc. of Image and Vision Computing, Auckland, New Zealand, 2000*. [http://projet-renault.enpc.org/papers\\_pdf/Background%20Subtraction%20:%20Tracking/ivcnz00.pdf](http://projet-renault.enpc.org/papers_pdf/Background%20Subtraction%20:%20Tracking/ivcnz00.pdf) (accessed August, 2008).
- Piccardi, M. (2004, Oct.). Background subtraction techniques: a review. In *IEEE International Conference on Systems, Man and Cybernetics, 2004*, Volume 4, pp. 3099–3104 vol.4. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed September, 2009).
- Rabaud, V. and S. Belongie (2006). Counting crowded moving objects. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, Volume 1, New York, US, pp. 705–711. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed February, 2007).
- Ronfard, R. (1994). Region-based strategies for active contour models. *International Journal of Computer Vision* 13(2), 229–251.
- Rubner, Y., C. Tomasi, and L. J. Guibas (2000). The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision* 40(2), 99–121. <http://ai.stanford.edu/~rubner/papers/rubnerIjcv00.pdf> (accessed January 2009).
- Salari, V. and I. K. Sethi (1990, Jan). Feature point correspondence in the presence of occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1), 87–91. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed August, 2008).
- Sato, K. and J. K. Aggarwal (2004). Temporal spatio-velocity transform and its application to tracking and interaction. *Computer Vision and Image Understanding*. 96(2), 100–128.
- Sharma, A. (2000). Crowd-behavior prediction using subjective factor based multi-agent system. In *IEEE International Conference on Systems, Man, and Cybernetics, 2000*, Volume 1, Nashville, TN, pp. 298–300. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed March, 2007).
- Shi, J. and C. Tomasi (1994, Jun). Good features to track. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. CVPR '94.*,

- 593–600. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed March, 2007).
- Sobel, I. and G. Fieldman (1968). A 3x3 isotropic gradient operator for image processing. presented at a talk at the Stanford Artificial Project in 1968, unpublished but often cited, orig. in *Pattern Classification and Scene Analysis*, Duda,R. and Hart,P., John Wiley and Sons,'73, pp271-2.
- Stauffer, C. and W. Grimson (2000, Aug). Learning patterns of activity using real time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8), 747–757. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed March 10, 2007).
- Streit, R. L. and T. E. Luginbuhl (1994). Maximum likelihood method for probabilistic multi-hypothesis tracking. In *Proceedings of the International Society for Optical Engineering (SPIE.)*, Volume 2235, pp. 394–405.
- Tao, H., H. S. Sawhney, and R. Kumar (2002, Jan). Object tracking with bayesian estimation of dynamic layer representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(1), 75–89.
- Torr, P. H. S. and A. Zisserman (1999). Feature based methods for structure and motion estimation. In *Proceedings of the International Workshop on Vision Algorithms ICCV '99*, London, UK, pp. 278–294. Springer-Verlag. ACM Digital Library. <http://portal.acm.org> (accessed July, 2008).
- Uncu, O., W. Gruver, D. Kotak, D. Sabaz, Z. Alibhai, and C. Ng (2006, 8-11). Gridbscan: Grid density-based spatial clustering of applications with noise. In *IEEE International Conference on Systems, Man and Cybernetics, 2006. SMC '06.*, Volume 4, pp. 2976–2981.
- Veenman, C. J., M. J. T. Reinders, and E. Backer (2001, Jan). Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(1), 54–72. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed August, 2008).
- Velastin, S. A., L. Khoudour, B. P. L. Lo, J. Sun, and M. A. Vicencio-Silva (2004). PRISMATICA: a multi-sensor surveillance system for public transport networks. In *12th IEE International Conference on Road Transport Information and Control, 2004.*



- RTIC 2004.*, London, UK, pp. 19–25. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed February, 2007).
- Wang, X., K. Tieu, and E. Grimson (2006). Learning semantic scene models by trajectory analysis. *European Conference on Computer Vision 2006. ECCV (3)* 3, 110–123. <http://www.informatik.uni-trier.de/~ley/db/conf/eccv/eccv2006-3.html> (accessed January, 2008).
- Yilmaz, A., O. Javed, and M. Shah (2006). Object tracking: A survey. *ACM Computing Surveys (CSUR)* 38(4), 13. ACM Digital Library. <http://portal.acm.org> (accessed January, 2008).
- Zhan, B., D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu (2008). Crowd analysis: a survey. *Machine Vision and Applications* 19(5-6), 345–357. ACM Digital Library. <http://portal.acm.org> (accessed July, 2009).
- Zhan, B., P. Remagnino, and S. A. Velastin (2005). Visual analysis of crowded pedestrian scenes. In *XLIII Congresso Annuale AICA 2005*, pp. 549–555. [http://dircweb.king.ac.uk/papers/Zhan\\_B.2005\\_399923/ZhanAICA2005.pdf](http://dircweb.king.ac.uk/papers/Zhan_B.2005_399923/ZhanAICA2005.pdf) (accessed June 15, 2009).
- Zhang, L., B. Curless, A. Hertzmann, and S. M. Seitz (2003). Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multi-view stereo. In *Proceedings. Ninth IEEE International Conference on Computer Vision, 2003.*, pp. 618–625 vol.1. IEEEXplore database. <http://ieeexplore.ieee.org> (accessed September, 2009).
- Zheng, H., Z. Wang, L. Zhang, and Q. Wang (2010, 27-29). Clustering algorithm based on characteristics of density distribution. In *2nd International Conference on Advanced Computer Control (ICACC), 2010*, Volume 2, pp. 431–435.
- Zivkovic, Z. and B. Krose (2004, July). An em-like algorithm for color-histogram-based object tracking. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Volume 1, pp. 798–803.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been

omitted or incorrectly acknowledged.