Visual Analysis of Videos of Crowded Scenes

A Thesis

Submitted to the Faculty

of

Drexel University

by

Louis Aloysious Kratz III

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

April 2012

Acknowledgments

I would like to begin by thanking my advisor Ko Nishino. Without doubt I would not have come this far without his advice and mentorship. His disciplined approach to the sciences has changed the way I approach problems and face hardships. In addition, his passion for research and thoughtful insights have both inspired and amazed me.

I would also like to thank my mentor, Dan Morris, and colleague, Scott Saponas at Microsoft Research for their inspiration and professional guidance. Without their encouragement, I doubt I would have the confidence to take true risks in my career.

Of course, my graduate career was not without difficulty. I would like to thank the multitude of individuals at Drexel University who supported and encouraged me during difficult times. Specifically: Geoff Oxholm, Prabin Bariya, Linge Bai, Peter Bogunovich, Julie Fisher, Andrea Negro, Amanda Kennedy, Kevin Lynch, and Bill Mongan.

Finally, I would like to thank my fiancée Hilary. Without her I would not be the person I am today, nor strive to be the person I want to be.

Table of Contents

List of Tables

List of Figures

Abstract

Visual Analysis of Videos of Crowded Scenes
Louis Aloysious Kratz III
Ko Nishino, Ph.D.

Automatic, vision-based analysis of crowds has implications in a number of fields, but faces unique challenges due to the large number of pedestrians within the scenes. The movement of each pedestrian contributes to the overall *crowd motion* (i.e., the collective motions of the scene's constituents) that varies spatially across the frame and temporally over the video. This thesis explores how to model the dynamically varying crowd motion, and how to leverage it to perform vision-based analysis on videos of crowded scenes. The crowd motion serves as a *scene-centric* constraint (i.e., representing the motion in the entire video), compared with conventional object-centric methods that build on individual constituents. By exploring what information the crowd motion can represent, we demonstrate the impact of leveraging our model on three problems facing video analysis of crowded scenes.

First, we represent the crowd motion using a novel statistical model of *local motion patterns* (i.e., the motion in local space-time areas). By doing so, we may learn the spatially and temporally varying underlying structure of the crowd motion from an example video of crowd behavior.

Second, we use our model to represent the typical crowd activity (i.e., the crowd's *steady-state*) and detect unusual events in local areas of the video. Specifically, we identify local motion patterns that statistically deviate from our learned model. Our space-time model enables detection and isolation of unusual events that are specific to the scene and the location within the video.

Next, we use the crowd motion as an indicator of an individual's motion to perform tracking. Specifically, we predict the local motion patterns at different space-time locations of the video and use them as a prior to track individuals in a Bayesian framework. Leveraging the crowd motion provides an accurate prior that dynamically adapts to the space-time variations of the crowd.

Finally, we explore how to measure how much individual pedestrians conform to the movement of the crowd. To achieve this, we use our crowd model to indicate the future locations of pedestrians, and compare the direction they would move to their instantaneous optical flow. By identifying deviations from the crowd, we identify global unusual events and augment our tracking method to model the individuality of each target.

We compare with conventional object-centric methods and those that do not encode the space-

time varying motion of the crowd. We demonstrate that our scene-centric approach (i.e, one that starts with the crowd motion) advances video analysis closer to the robustness and dependability needed for real-world video analysis of scenes containing a large number of pedestrians.

Chapter 1: Introduction

1.1 Background and Motivation

Vision-based analysis of videos of crowded scenes, such as those shown in Fig. 1.1, has the potential to impact a wide number of fields, but computer vision methods have been hindered by the unique challenges associated with crowded scenes.

Video analysis of crowds has vast implications in video surveillance, event planning, and disaster avoidance. The prevalence of video surveillance systems has created a dire need for vision-based methods that can assist or replace human operators. Many surveillance operators are tasked with watching multiple screens simultaneously, and the large number of pedestrians in even a single crowded scene is difficult to monitor. Vision methods also have the potential to identify and measure congestion, enabling event planners to anticipate the movement of pedestrians and adjust the environment to maximize crowd flow. Finally, crowded areas pose a high risk for disasters as a result of over-crowding, panic, or stampedes. By understanding the behavior of crowds, vision methods have the potential to automatically detect or anticipate crowd disasters.

Despite the need for video analysis, the complexity of crowded scenes pose unique challenges to vision-based methods. As shown in Fig. 1.1, the large number of pedestrians in crowded scenes results in frequent partial or complete occlusions. In addition, almost every individual is completely surrounded by other moving pedestrians. The scenes are so dense, in fact, that identifying foreground pixels is futile as the background is rarely visible. Finally, pedestrians may move in any number of different directions depending on their personal goals, neighboring pedestrians, and the physical obstacles within the scene, resulting in a heterogeneous and dynamically evolving crowd motion that is often too complex for conventional methods.

A promising way to handle the complexity of crowded scenes is through learning. Surveillance cameras are often fixed and operate continuously, resulting a large amount of video of typical behavior. Pedestrians move through the scene depending on their personal goals, but their motion is

Figure 1.1: Example frames of crowded scenes.

also influenced by *scene-specific* factors such as the environment, the crowd's density [68], or even their culture [17]. Since these scene-specific factors change gradually, if at all, the collective motion of pedestrians within the scene (i.e., the *crowd motion*) tends to repeat and can be learned from example videos of the crowd. This learned model can then be used as an estimate of the crowd behavior in a video of the same scene recorded at a different time.

Conventional video analysis methods learn the behavior of the scene in three steps: detecting constituents, tracking constituents, and compiling the tracking results into higher order models. The applicability of such object-centric methods is limited to scenes with relatively few constituents. Discerning individuals in crowded scenes is difficult since they are typically surrounded by other moving pedestrians. Tracking is also difficult due to the frequent partial or complete occlusions in crowded scenes. Finally, such methods suffer from problems of scale: each new pedestrian that enters the scene increases the complexity of the model, making them intractable on very crowded scenes.

Rather than learning the crowd motion from the trajectories of each pedestrian, a more tractable approach is to treat the crowd holistically. Holistic methods model relationships between low-level motion estimates to represent the motion of the entire scene. Often, these estimates represent the motion in small, local space-time areas of the video (i.e., a *local motion pattern*). The challenge faced by holistic techniques is to accurately, efficiently, and faithfully represent the crowd motion. The crowd motion is complex as it can vary spatially across the frame and temporally over the

video. By learning these spatio-temporal variations, however, the crowd motion may be used as a scene-centric constraint to analyze videos of crowded scenes.

1.2 Related Work

Crowds have received attention from researchers in physics, sociology, graphics, and computer vision. While vision researchers aim to understand the behavior of the crowd (or, in some cases, a pedestrian within the crowd) from visual evidence, work in physics and graphics aim to analytically model the crowd to create accurate simulation models. Often, such simulations are used to evaluate evacuation dynamics [66] or public space design in order to avoid crowd disasters. We discuss related vision work and refer to [82, 80] for a full discussion of the relationship between simulation and vision.

1.2.1 Object-Centric Video Analysis

Conventional methods for video analysis are object-centric (i.e., begin by analyzing each of the scene's constituents). Such methods detect the scene's constituents (most frequently pedestrians or automobiles), track them, and then analyze the trajectories to model the behavior of the constituents. These methods work well on scenes that are relatively sparse (roughly 5 to 20 pedestrians) and, as noted by Zhan [79], are not appropriate for very crowded scenes. We review related object-centric work that are designed for crowded scenes, but emphasize that they still have challenges in videos containing high density crowds.

Detecting the scene's constituents is often the first step in object-centric video analysis. Zhao et al. [81], for example, track pedestrians in videos of crowds by detecting each individual using a model of human shapes. Rodriguez and Shah [65] detect pedestrians using a voting scheme on the contours around each individual. The contours are computing by subtracting the background from each video frame. In high density crowded scenes, however, the background is rarely visible and pedestrians are often partially occluded, making the contours difficult to estimate. Leibe et al. [47] also segment pedestrians from the background, but use global image cues to add robustness partial occlusions. Their method handles some partial occlusions well, but assumes that the torso of the pedestrian is visible. This is often not the case in near-view scenes, such as those shown in Fig. 1.1,

where only the heads of most pedestrians are visible.

Other work detect pedestrians by assuming that they exhibit unique motion. Browstow et al. [14] group short feature tracks (or "tracklets") to identify similarly moving pedestrians. They assume that the subjects move in distinct directions and thus disregard possible local motion inconsistencies between different body parts. As noted by Okabe et al. [69], such inconsistencies cause a single pedestrian to be detected as multiple targets. In addition, pedestrians that are moving in the same direction are identified as a single group. Crowded scenes, especially when captured in relatively near-field views as is often the case in video surveillance, necessitate a method that represents the multiple motions of a single individual or similar motions of different pedestrians.

After detection, the constituents are tracked as they move through the scene. Data association methods, such as that of Betke et al. [8] or Gilbert and Bowden [23], track multiple targets in cluttered scenes by associating detection results of consecutive frames. These techniques assume that the detection is always reliable, and thus degrade in very crowded scenes. Wu and Nevatia [77] are able to track partially occluded pedestrians by detecting body parts, rather than the full pedestrian. The data association problem itself is NP-Hard, and thus becomes less tractable in scenes with a large number of pedestrians. Often, approximation techniques are used to estimate a solution such as the Bayesian framework of Li et al. [50].

Other data association methods do not rely on detecting individuals. Khan et al. [39] model the interaction among detected interest points to improve the tracking of each object. Hu et al. [35] use a Markovian model on each tracked point to augment data association in generic domains. As noted by Khan et al. [40], however, a single point may be shared between multiple targets and result in ambiguities. Shared points are often the result of motion boundaries or clutter, both of which occur frequently in videos of crowded scenes.

After tracking, the trajectories are used to characterize behaviors about the constituents within the scene. Wang et al. [75], for example, cluster trajectories to learn the common routes taken by pedestrians and automobiles. Dee and Hogg [22] use the tracking information to identify pedestrians that deviate from a goal-specific behavior. Hu et al. [34] learn global motion patterns (i.e., that

describe motion over the entire frame) and use them to detect anomalies and predict future behaviors. Johnson and Hogg [37] estimate different distributions of trajectories, and attach semantics to each in order to identify specific events within the scene. Such methods not only depend on reliable detection and tracking, which may not be available in videos of crowded scenes, but also face problems of scale. As more pedestrians enter the scene, the complexity of these methods increases and may become intractable with even moderately dense crowds.

1.2.2 Local Motion Patterns

To address the complexity of real-world scenes, many researchers propose *holistic* techniques that characterize the scene as a collection of local motion estimates (which we refer to as *local motion patterns*), rather than a collection of constituents. Often, holistic methods aim to identify behaviors within the scene that are part of the same physical process [32]. For example, Yang et al. [78] use a "bag-of-words" model to identify and detect common automobile traffic patterns. These can then be used to mine events [48] or to segment the video [49]. A key challenge is how to represent the local motion pattern while maintaining a computationally feasible model. Here, we review related work in this area, but note that most methods in this section are not intended for use in crowded scenes.

Yang et al. [78] use the 2D frame location and motion direction to represent local motion patterns. Similar to Wang et al. [74] and Hospedales et al. [31], the motion direction is quantized into one of four primary directions. Doing so makes the bag-of-words model discrete (and thus computationally efficient), but limits the motion of constituents to relatively rigid movements. As such, these methods are suitable for scenes where the motion of constituents is highly structured, such as automobile intersections, but are not suitable for crowded scenes of pedestrians.

Other representations have been useful for action detection. Key et al. [38] detect actions in cluttered scenes by comparing gram matrices of spatio-temporal gradients. They use the motion-similarity measure from Boiman and Irani [67] that assumes each local area contains motion in a single direction. In real-world crowded scenes, however, the near-proximity of pedestrians leads to frequent motion boundaries whose optical flow is undefined. In high-density heterogeneous crowds,

this problem intensifies and a single motion direction may not accurately describe the motion in many instances.

Histograms of oriented gradients (HoG) features have also been used to describe space-time volumes for human detection [21] and action recognition [46]. The HoG feature is computed on the spatial gradients (the temporal gradient is not used), though they have been extended to the 3D space-time gradient [41]. The orientation of spatial gradients encodes the structure of the pedestrian's appearance, and thus is not suitable when only motion is necessary.

The relationship between local motion patterns has also been used to improve the tracking of motion boundaries. Nestares and Fleet [58] increase the continuity of the motion boundary tracking from Black and Fleet [10] by using motion patterns from the local area surrounding the tracked region. They assume that spatially neighboring motion patterns are similar, which is not the case in heterogeneous or dynamically evolving crowds.

1.2.3 Crowd Models

Many work have borrowed ideas from holistic video analysis to address the challenges in videos of crowds. Such methods are similar in spirit to ours, but often make overly simplistic assumptions regarding the motion of pedestrians or the crowd itself. We further demonstrate the difference between, and compare the results of, our model and these methods in the appropriate sections of this dissertation.

Mahadevan et al. [51] describe the typical dynamics of the crowd with a mixture of dynamic textures (previously used for segmentation by Chan et al. [15]). Using dynamic textures, however, retains appearance variations which can introduce noise into the model and degrade results.

Shah et al. [57, 55, 2] model crowds based on a hydrodynamics model that essentially treats each pedestrian as a particle in a fluid. As noted by Still [68], however, specific behaviors that occur in crowds, such as lane formations or clustering, do not occur in fluids. While particles are affected only by the *external* forces around them (such as other particles or the environment), the motion of pedestrians is a result of *both* external forces and their individual desires. Such differences between individual pedestrians form dynamic space-time structures in the crowd motion that can not be

represented with a hydrodynamics model.

Other work assumes that the crowd flow is constant over the entire video. Ali and Shah [1] average the optical flow over a video clip, and use it to model a Finite Time Lyapunov Exponent field for segmenting the motion of the crowd. Similarly, Mehran et al. [56] measure the "social force" [26] by comparing the instantaneous optical flow to the optical flow averaged over the video clip. Raghavendra et al. [61] also estimate the social force, but do so using a particle swarm method that clusters similar motion vectors. In many crowded scenes, especially those with unconstrained environments, the motion of pedestrians can change dramatically in a short period of time as individuals move towards different goals.

Some work assume the crowd exhibits homogeneous motion in each area of the scene. Hu and Shah [33], for example, identify global motion patterns (i.e., ones that take up the entire frame) in crowded scenes by clustering optical flow vectors in similar spatial regions. Similarly, Cheriyadat and Radke [19] detect dominant motions in crowds by clustering low-level tracked features. Such methods can not handle dynamically varying crowds or those with heterogeneous motions in local areas. A crosswalk, for example, naturally has pedestrians moving in two directions who emerge together as they pass each other.

Other methods capture the multi-modal nature of the crowd, but ignore the important temporal relationship between sequentially occurring motions exhibited by pedestrians. Rodriguez et al. [62] use a topical model (similar to the bag-of-word models) over quantized optical flow directions to describe the crowd motion. They later improve their tracking using a crowd density estimate [63]. Though they model the heterogeneous nature of the crowd, they do not encode the relationship between temporally co-occurring motions. By disregarding the temporal variations in the motions exhibited by pedestrians, these approaches cannot represent the underlying temporal pattern within the crowd motion.

Andrade et al. [3, 4] captures the temporal structure of the crowd by training hidden Markov models on optical flow vectors. They demonstrate that their method is a good indicator of emergency situations in simulated crowd flow data. Real-world crowded scenes, however, were not evaluated.

1.3 Overview

In this dissertation, we demonstrate how a space-time model of the crowd motion may be learned from a video of typical crowd activity and used as a scene-specific constraint to analyze videos of crowded scenes. Specifically, we explore what information can be leveraged from the crowd model to address three challenges in the video analysis of crowded scenes. Our contributions are organized as follows:

Space-Time Model of Crowd Motion In Chapter 2, we present a novel space-time statistical model for representing the underlying structured pattern of the crowd motion. We start by representing the motion in each local space-time region (i.e., the *local motion pattern*) with a directional distribution of optical flow. By doing so, we encode the possibly complex motions pedestrians can exhibit within crowded scenes by the uncertainty in the optical flow. Next, we train a collection of hidden Markov models (HMMs) over the local motion patterns in order to learn the multi-modal and dynamically varying crowd motion. The collection of HMMs represents the spatially and temporally varying motions that can occur in crowded scenes, and does so without identifying or tracking each pedestrian.

Anomaly Detection In Chapter 3, we train our model of the crowd motion on videos of typical crowd behavior to represent the *steady-state* or "usual" motion of the crowd. We then detect local unusual events in videos of crowded scenes by identifying local motion patterns that statistically deviate from the learned model. Since our model is trained on a video of the same scene, the detected unusual events are *scene-specific*, meaning that our method may be tailored to each specific environment. In addition, our space-time model allows us to detect unusual events that are specific to the location within the frame.

Tracking In Chapter 4, we leverage the crowd motion to estimate the movement of each individual within the scene. Specifically, we use our model as a space-time prior for tracking individuals through videos of crowded scenes. We predict the local motion pattern that describes the movements of pedestrians at each space-time location of the video. We then use this prediction

as a prior on the state-transition distribution of a particle filter to track the individual. Using our unique model yields a prior that changes spatially and temporally over the video, enabling accurate tracking of individuals moving through the dynamically changing scene.

Conformance To The Crowd In Chapter 5, we leverage our model to explore the relationship between the individual and the crowd. Specifically, we measure how much each pedestrian conforms to the flow of the crowd. To achieve this, we use the crowd motion to estimate the future locations of pedestrians, and treat this as an indicator of where they intend to move. We then compare this to the instantaneous optical flow to measure the conformance of the pedestrian to the crowd. Low conformance indicates that a pedestrian is deviating from the crowd, often causing crowd-based methods to degrade. In such cases, we demonstrate that conformance may be used to dynamically adjust our motion prior for tracking and as an indicator of unusual events.

Finally, in Chapter 6, we summarize our contributions and discuss future work.

Chapter 2: Crowd Motion Model

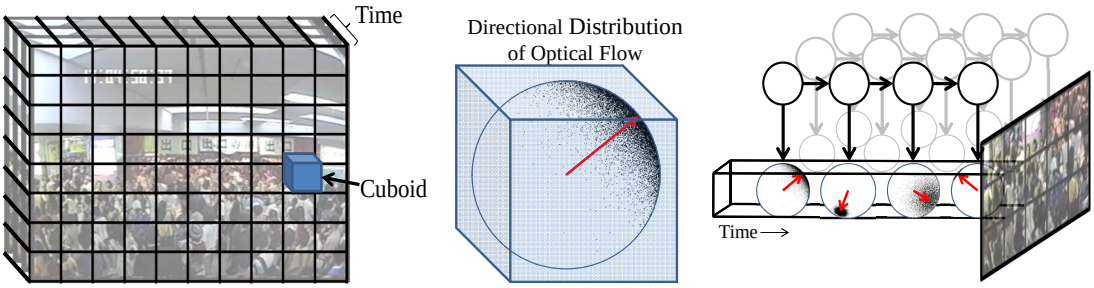
2.1 Introduction

We begin by representing the crowd motion with a spatially and temporally varying statistical model. In the absence of other pedestrians, individuals move in straight lines towards their destinations. In higher densities, however, they naturally form organized structures (i.e., emergent behaviors) to utilize the available space and achieve a higher flow [44]. Such behaviors form visual structures (such as lanes or clusters) that can be identified as global motion patterns by grouping trajectories or optical flow vectors [32]. The motion of pedestrians within these structures, however, may still vary significantly and introduce noise. To overcome these challenges, we model the crowd motion as a dynamically evolving structure of local motion patterns, eliminating the need to track each pedestrian while efficiently representing their various motions.

Fig. 2.1 shows an overview of our model. First, as shown in Fig. 2.1(a), we divide a training video into spatio-temporal sub-volumes, or "cuboids," defined by a regular grid. Second, as shown in Fig. 2.1(b), we model the motion of pedestrians through each cuboid (i.e., the local motion pattern) with a 3D directional distribution of optical flow. Next, as shown in Fig. 2.1(c), we train a hidden Markov model (HMM) over the local motion patterns at each spatial grid location. The hidden states of the HMMs encode the multiple possible motions that can occur at each spatial location. The transition probabilities of the HMMs encode the time-varying dynamics of the crowd motion. We represent the crowd motion by the collection of HMMs, encoding the spatially and temporally varying motions of pedestrians that comprise the entire crowd motion.

2.2 Comparison with Related Work

Here, we compare our model with other methods that are designed specifically for crowded scenes. Our model has three unique characteristics that distinguish it from over other methods. First, our model is varies in *space* as well as *time*. Second, we model the crowd motion by starting with local



(a) Training Video (b) Local Motion Pattern (c) Collection of HMMs

Figure 2.1: An overview of our model of the crowd motion.

motion patterns. This enables the model to scale with the modality of different crowd behaviors, rather than the number of pedestrians. Finally, since our model is based on hidden Markov models, it may be learned from an example video of the scene and used to analyze videos of the same scene recoded at a different time.

Ali and Shah [1, 2] use "particle advection" to analyze the motion of the crowd. Specifically, they average the optical flow over a video clip, and advance particles through the flow fields to measure the crowd dynamics. Their approach assumes that the crowd does not change over time. In contrast, the hidden states in our collection of HMMs represent the multi-modal motions that exist in videos of crowded scenes. In addition, Ali and Shah analyze the crowd in a single video clip. We train our model on an example of typical crowd behavior, and then use it to analyze videos of the same scene recorded at a different time.

Other work represent the multi-modal motions that may occur within the scene, but do not encode the temporal relationship between them. Mehran et al. [55] use "streaklines," a concept from hydrodynamics, to segment videos of crowds and track pedestrians. Though streaklines encode some temporal information, Mehran et al. do not model the dynamics between consecutive streaklines motion of the crowd. Similarly, the correlated topical model from Rodriguez et al. [62, 64] encodes the multi-modal motions, but disregards the relationship between motions that co-occur.

Other work estimate global motion patterns from videos of crowds. Hu and Shah [33] cluster optical flow vectors in similar spatial regions to estimate motion patterns. Such work is applicable

to scenes where the motion of the crowd has large, stable patches of heterogeneous flow. In contrast, we represent the motion in small, space-time areas with a *local* motion pattern and capture the dynamically varying heterogeneous crowd with a collection of HMMs.

Andrade et al. [3, 4] also use a collection of hidden Markov models to represent the crowd motion. The observations to the HMM are vectors of pixel locations and optical flow estimates. While these may be viewed as a form of local motion patterns, they do not directly encode the variability in motion that can occur due to poor texture or aperture problems. In contrast, our directional distributions of optical flow directly encode the uncertainty in the optical flow estimate as we later demonstrate.

2.3 Local Motion Patterns

First we seek to represent the motion in each cuboid in the video volume, i.e., the local motion patterns. The optical flow can be reliably estimated when the cuboid contains good texture and motion in a single direction with constant velocity. The motion in cuboids from real-world crowded scenes, however, may be difficult to estimate reliably. A cuboid may contain *complex* motion, i.e., motion exhibited by two objects moving in multiple directions or a single object that changes direction or speed. In addition, cuboids may contain little or no texture and have indeterminable motion due to the aperture problem. To handle these different cases, we model each local motion pattern with a directional distribution of optical flow whose variance encodes the uncertainty in the motion estimate. To achieve this, we first discuss the relationship between the spatio-temporal gradients and the motion within the cuboid. We then use the characteristics of the spatio-temporal gradients to estimate a distribution of optical flow vectors.

Spatio-Temporal Gradients

Let $\nabla I(x, y, f)$ be a 3×1 vector of image gradients estimated in the horizontal, vertical, and temporal directions, respectively, at 2D pixel location (x, y) and frame f. Horn and Schunck [30] show the relationship between the spatio-temporal gradient and the optical flow $\mathbf{q} = [u, v, 1]^T$ by the constraint

$$\nabla I(x, y, f)^T \mathbf{q} = 0. \tag{2.1}$$

Since $\mathbf{q}$ has two unknowns, the problem of estimating $\mathbf{q}$ from a single gradient estimate is ill-posed.

Often, it is assumed that the flow is constant in the space-time area around (x, y, f), and surrounding gradients are used to estimate the optical flow $\mathbf{q}$. Let $\{\nabla I_i | i = 1 \dots N\}$ be a set of N spatio-temporal gradients (we have dropped x, y, f for notational convenience) computed at the different pixel locations of the cuboid. Assuming that the flow over all N points is constant, then

$$\begin{bmatrix} \nabla I_1^T \\ \vdots \\ \nabla I_i^T \\ \vdots \\ \nabla I_N^T \end{bmatrix} \mathbf{q} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}^T . \tag{2.2}$$

Multiplying both sides by $[\nabla I_1, \dots, \nabla I_N]$ yields the more compact form

$$\sum_i^N \nabla I_i \nabla I_i^T \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} . \tag{2.3}$$

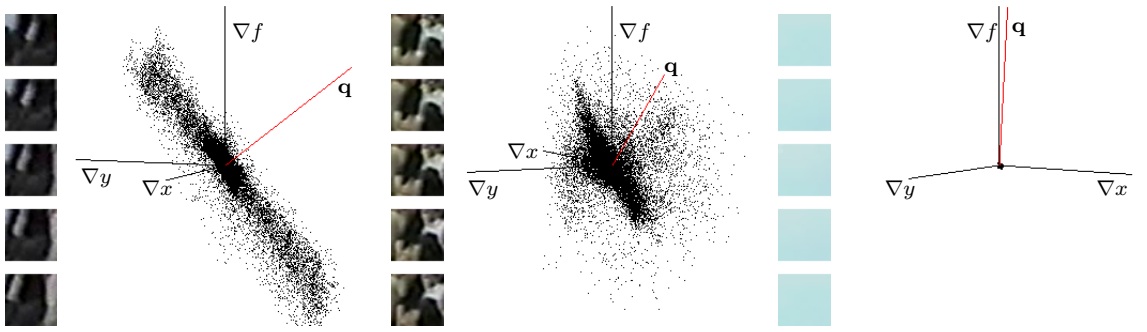
The 3×3 matrix

$$\mathbf{G} = \sum_i^N \nabla I_i \nabla I_i^T, \tag{2.4}$$

is known as the structure tensor [76].

The solution to Eq. 2.3 (i.e., the estimate of the optical flow $\mathbf{q}$) is the eigenvector $\tilde{\mathbf{q}} = [\tilde{q}_x, \tilde{q}_y, \tilde{q}_t]^T$ of $\mathbf{G}$ with the smallest eigenvalue [76]. Assuming the change in time is 1,

$$\mathbf{q} = \left[\frac{\tilde{q}_x}{\tilde{q}_t}, \frac{\tilde{q}_y}{\tilde{q}_t}, 1 \right]^T . \tag{2.5}$$



(a) Uniform Motion (b) Complex Motion (c) No Motion

Figure 2.2: The shape of space-time gradients characterizes the motion within the cuboid.

Fig. 2.2 shows the spatio-temporal gradients for three different cuboids: one with good texture and uniform flow, one with complex motion, and one with poor texture. The shape of a cuboid's spatio-temporal gradients characterizes the motion within the cuboid. As shown in Fig. 2.2(a), cuboids containing motion in a single direction have spatio-temporal gradients that are coplanar. This plane is orthogonal to the 3D optical flow $\mathbf{q}$ [76] since the dot product between $\mathbf{q}$ and any gradient is 0 as stated in Eq. 2.1. Cuboids containing complex motion result in a spatio-temporal gradients that are not coplanar (Fig. 2.2(b)) and tend to exhibit a spherical or ellipsoidal shape. Finally, cuboids containing poor texture (Fig. 2.2(c)) have spatio-temporal gradients that lie near the origin.

In our previous work [42, 43] we represent each local motion patterns by a 3D Gaussian distribution of spatio-temporal gradients. Each local motion pattern is defined by the mean vector

$$\overline{\nabla I}_t = \frac{1}{N}\sum_{i}^{N} \nabla I_i \tag{2.6}$$

and a covariance matrix

$$\mathbf{\Sigma}_t = \frac{1}{N}\mathbf{G} - \overline{\nabla I}_t^T \overline{\nabla I}_t. \tag{2.7}$$

The 3D Gaussian representation encodes the motion within the cuboid by the shape of the spatio-temporal gradients, but retains appearance information in the form of the spatial gradients (i.e., the first and second dimensions of $\overline{\nabla I}$ and $\mathbf{\Sigma}$). As a result, the crowd motion model has a high

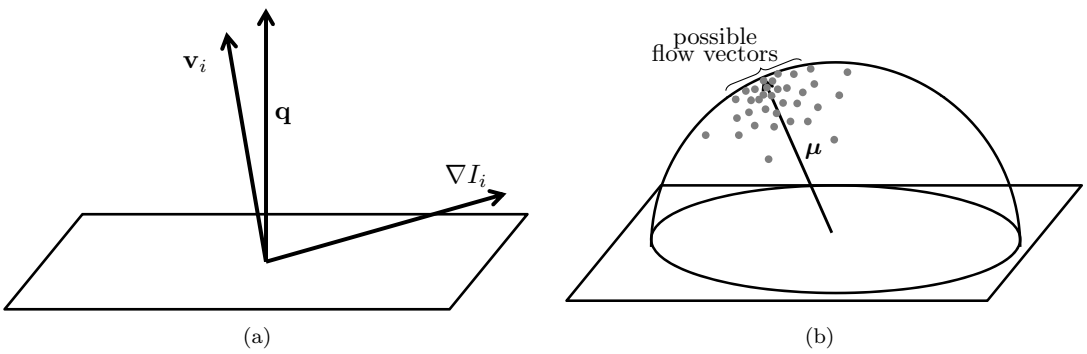


Figure 2.3: Illustration of estimating a distribution of optical flow vectors from spatio-temporal gradients.

dimensionality and requires longer training data to capture different appearance variations. We exploit this appearance information to aid in tracking [43], but note that removing appearance information results in a more compact and accurate model. Next, we present a more compact representation based on directional distributions of optical flow.

Directional Distributions

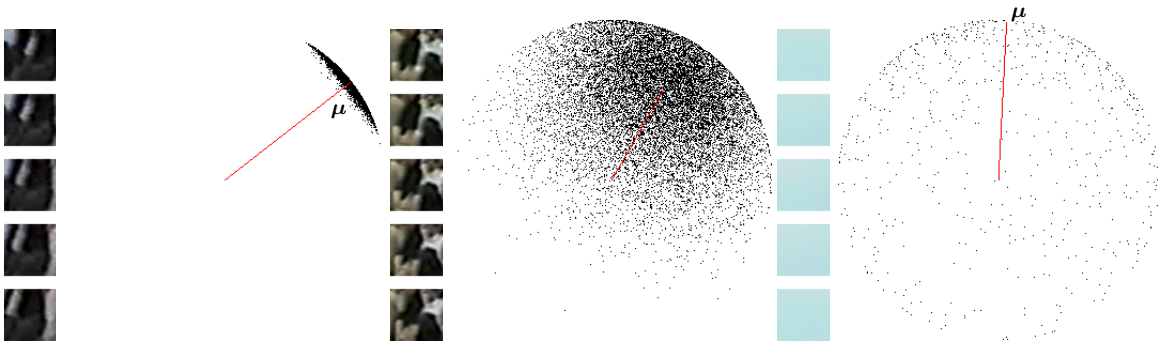
We use the shape of the spatio-temporal gradients to estimate a directional distribution of 3D optical flow vectors that describe the possible motions that can occur inside of the cuboid. As illustrated in Fig. 2.3, we consider a spatio-temporal gradient ∇I_i that is not on the plane defined by the 3D optical flow vector $\mathbf{q}$. Such a point suggests that the actual motion within the cuboid may be in another direction

$$\mathbf{v}_i = \frac{\nabla I_i \times \mathbf{q} \times \nabla I_i}{|\nabla I_i \times \mathbf{q} \times \nabla I_i|} \tag{2.8}$$

where $\times$ is the cross-product. Note that $\mathbf{v}_i$ is orthogonal to ∇I_i, and thus satisfies the optical flow constraint in Eq. 2.1 for ∇I_i.

As shown in Fig. 2.3(b), the distribution $\{\mathbf{v}_i \,|\, i{=}1, \ldots, N\}$ exists on the unit sphere. A natural representation is the von Mises-Fisher distribution [54]

$$\mathrm{p}(\mathbf{v}) = \frac{1}{c(\kappa)} \exp\left\{\kappa \boldsymbol{\mu}^T \mathbf{v}\right\} , \tag{2.9}$$



(a) Uniform Motion (b) Complex Motion (c) No Motion

Figure 2.4: The variance of optical flow characterizes the motion with each cuboid.

where $\boldsymbol{\mu}$ is the mean direction, $c(\kappa)$ is a normalization constant, and $\kappa > 0$ is a real number known as the concentration parameter.

We represent the local motion pattern within each cuboid by estimating a a von Mises-Fisher distribution to the possible optical flow vectors $\{\mathbf{v}_i \,|\, i=1,\ldots,N\}$. Mardia and Jupp [54] show that the sufficient statistic for estimating $\boldsymbol{\mu}$ and κ is

$$\mathbf{r} = \frac{1}{N}\sum_{i}^{N}\mathbf{v}_i \ , \tag{2.10}$$

and thus

$$\boldsymbol{\mu} = \frac{1}{|\mathbf{r}|}\mathbf{r} \ . \tag{2.11}$$

They also show that

$$\kappa = \begin{cases} \dfrac{1}{(1-|\mathbf{r}|)} & \text{if } |\mathbf{r}| \geq 0.9 \\ \\ A_3^{-1}(|\mathbf{r}|) & \text{otherwise,} \end{cases} \tag{2.12}$$

where

$$A_3(|\mathbf{r}|) = \coth(|\mathbf{r}|) - \frac{1}{|\mathbf{r}|} \ . \tag{2.13}$$

In our implementation, we use the tabulated data from Mardia and Jupp [54] to compute $A_3^{-1}(|\mathbf{r}|)$.

As illustrated in Fig. 2.4, the concentration parameter κ characterizes the uncertainty in the optical flow estimate within the cuboid. Cuboids containing motion in a single direction have a

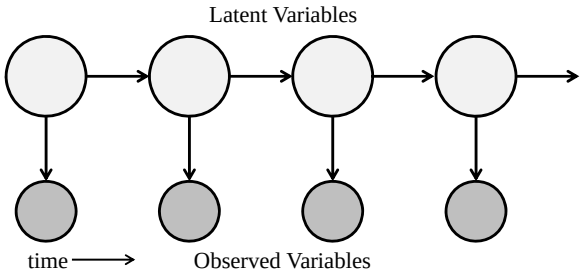


Figure 2.5: Graphical model of the hidden Markov model.

high concentration parameter, yielding a narrow distribution. Cuboids with complex motion have a wide distribution, indicating that motion may occur in different directions. Cuboids with little or no texture have distributions across the entire sphere, indicating that motion may be occurring in any direction.

In summary, each local spatio-temporal motion pattern O is defined by a mean 3D optical flow vector $\boldsymbol{\mu}$ and a concentration parameter κ that encodes the uncertainty of the estimate.

2.4 HMMs of Local Motion Patterns

Now that we have a representation of the local motion patterns, we model a collection of hidden Markov models (HMMs) to represent the spatially and temporally varying crowd motion.

2.4.1 Overview of Hidden Markov Models

We briefly introduce hidden Markov models (HMMs) to provide the necessary background for future sections. We refer to Rabiner [60] for a complete introduction.

Fig. 2.5 shows the graphical model of an HMM. The shaded circles represent the observed variables, and the white circles represent the latent variables. The latent variables $\{s_t|t = 1, \ldots, T\}$ of the HMM are assumed to follow the Markov property: each latent variable depends on the previous. We only consider first degree models, where each latent state is dependent on the single previous one. Given the latent variables, the observed variables $\{O_t|t = 1, \ldots, T\}$ are conditionally independent.

In HMMs, the latent variables are discrete, taking on one of J values. Each HMM is defined by a $J\times1$ initial state probability vector $\boldsymbol{\pi}$, a $J\times J$ state transition matrix $\mathbf{A}$, and the emissions densities $\{\mathrm{p}(O_t|s_t=j)\,|\,j=1,\ldots,J\}$. The likelihood of starting in a specific state j is encoded by the initial

probability vector

$$\boldsymbol{\pi}(j) = \mathrm{p}(s_1 = j). \tag{2.14}$$

The likelihood of transitioning from state i to state j is represented by the state transition matrix

$$\mathbf{A}(i,j) = \mathrm{p}(s_{t+1} = j | s_t = i) \ . \tag{2.15}$$

The emission densities can take on various forms, and we discuss our formulation in Sec. 2.4.2.

A key problem in HMMs is to compute the likelihood of an observation sequence $\{O_1, \ldots, O_T\}$. This is achieved efficiently using dynamic programming by the Forwards-Backwards algorithm [60]. We review the forwards step here, as we use it extensively in this work. Let $\mathbf{b}_t$ be a $J \times 1$ vector of likelihoods where

$$\mathbf{b}_t(j) = \mathrm{p}(O_t | s_t = j) \ . \tag{2.16}$$

In the forwards step, dynamic programming is used to compute the message

$$\boldsymbol{\alpha}_t(j) = \mathrm{p}(s_t = j, O_1, \ldots, O_t) \ , \tag{2.17}$$

where $O_1, \ldots, O_t$ are the observations up to time t. After the first observation, the message is initialized

$$\boldsymbol{\alpha}_1(j) = \mathbf{b}_1(j)\boldsymbol{\pi}(j). \tag{2.18}$$

Subsequent messages are computed by the update

$$\boldsymbol{\alpha}_t(j) = \mathbf{b}_t(j) \sum_i^J \boldsymbol{\alpha}_{t-1}(i)\mathbf{A}(i,j). \tag{2.19}$$

Often $\boldsymbol{\alpha}_t$ is scaled by it's magnitude after each update to avoid numerical problems. This yields the posterior

$$\hat{\boldsymbol{\alpha}}_t(j) = \frac{\boldsymbol{\alpha}_t(j)}{|\boldsymbol{\alpha}_t|} = \mathrm{p}(s_t = j | O_1, \ldots, O_t) \ . \tag{2.20}$$

After the backwards step of the Forwards-Backwards algorithm, we may compute the full posterior

$$\boldsymbol{\gamma}_t(j) = \mathrm{p}(s_t = j | O_1, \ldots, O_T) \tag{2.21}$$

which is used during training to update the parameters of the HMM.

An important aspect of the forwards step is that it may be computed online. When each new observation O_t becomes available, the new posterior $\hat{\boldsymbol{\alpha}}_t$ may be computed efficiency by Eq. 2.19. We use this characteristic of HMMs in our applications to retain the possibility of online operation.

2.4.2 Emission Densities Over Local Motion Patterns

Next, we turn our attention to the form of the HMM's emission densities $\{\mathrm{p}(O_t|s_t = j) \,|\, j = 1, \ldots, J\}$. Each observation $O_t = \{\boldsymbol{\mu}_t, \kappa_t\}$ is a local motion pattern, defined by a 3D mean optical flow vector $\boldsymbol{\mu}_t$ and concentration parameter κ_t. Often complex observations are quantized using a codebook, making the emission densities discrete. This can decrease the training time, but reduces the amount of information represented by each emission density.

Rather than quantizing our local motion patterns, we analytically model the emission densities by imposing priors over $\boldsymbol{\mu}_t$ and κ_t. To achieve this, we consider $\boldsymbol{\mu}_t$ and κ_t statistically independent such that

$$\mathrm{p}(O_t|s_t = j) = \mathrm{p}(\boldsymbol{\mu}_t|s_t = j)\mathrm{p}(\kappa_t|s_t = j) \ . \tag{2.22}$$

We model $\mathrm{p}(\kappa_t|s_t = j)$ as a Gamma distribution defined by a shape parameter a^j and scale parameter θ^j. We model $\mathrm{p}(\boldsymbol{\mu}_t|s_t = j)$ as a von-Mises Fisher distribution (i.e., the conjugate prior on $\boldsymbol{\mu}_t$ [53]) defined by a mean direction $\boldsymbol{\mu}_0^j$ and concentration parameter κ_0^j . We provide evidence that $\boldsymbol{\mu}_t$ and κ_t follow these distributions in Sec. 2.5.2.

Training the HMM's is achieved using the Baum-Welch [60] algorithm. During training, the posteriors $\boldsymbol{\gamma}_t$ from Eq. 2.21 are used to update the emission density parameters $\{\boldsymbol{\mu}_0^j, \kappa_0^j, a^j, \theta^j\}$. Updating $\boldsymbol{\mu}_0^j$ and κ_0^j is similar to fitting von Mises-Fisher distributions as discussed in Sec. 2.3. The

average observed vector

$$\mathbf{r}_0^j = \frac{\sum_t^T \boldsymbol{\gamma}_t(j)\boldsymbol{\mu}_t}{\sum_t^T \boldsymbol{\gamma}_t(j)} \tag{2.23}$$

is used to compute the mean direction

$$\boldsymbol{\mu}_0^j = \frac{1}{|\mathbf{r}_0^j|}\mathbf{r}_0^j \tag{2.24}$$

and concentration parameter

$$\kappa_0^j = \begin{cases} \dfrac{1}{(1-|\mathbf{r}_0^j|)} & \text{if } |\mathbf{r}_0^j| \geq 0.9 \\ \\ A_3^{-1}(|\mathbf{r}_0^j|) & \text{otherwise,} \end{cases} \tag{2.25}$$

as in Sec. 2.3. There is no closed-form solution to estimating a^j, and thus we use the numerical technique from Choi and Wette [20]. Given an estimate of a^j maximum likelihood is used to estimate the scale

$$\theta^j = \frac{\sum_t^T \boldsymbol{\gamma}_t(j)\kappa_t}{a^j \sum_t^T \boldsymbol{\gamma}_t(j)} \,. \tag{2.26}$$

As discussed in Sec. 2.3, our previous work [42, 43] defines the local motion pattern O_t by the parameters $\overline{\nabla I}_t$ and $\boldsymbol{\Sigma}_t$. We again emphasize that such a representation results in a higher dimensional model, but we review the formulation for completeness. Each emission density is

$$p\left(O_t | s_t = j\right) = \frac{1}{\sqrt{\pi\sigma_s^2}} \exp\left[\frac{-\tilde{d}\left(O_t, P_j\right)^2}{2\sigma_j^2}\right], \tag{2.27}$$

where P_j is the expected local motion pattern, σ_j is the standard deviation, and $\tilde{d}\left(\cdot\right)$ is the Kullback-Leibler (KL) divergence [45]. The expected motion pattern $P_j = \{\overline{\nabla I}_j, \boldsymbol{\Sigma}_j\}$ is computed during

(a) Concourse (b) Ticket Gate (c) Sidewalk (d) Intersection

(e) Platform (f) Escalator (g) Street (h) Synthetic

Figure 2.6: Frames from video of crowded scenes that we analyze in this dissertation.

training by

$$\overline{\nabla I}_j = \frac{\sum_t^T \gamma_t(j)\overline{\nabla I}_t}{\sum_t^T \gamma_t(j)}, \tag{2.28}$$

$$\mathbf{\Sigma}_j = -\overline{\nabla I}_j \overline{\nabla I}_j^T + \frac{1}{\sum_t^T \gamma_t(j)} \sum_t^T \gamma_t(j)(\mathbf{\Sigma}_t + \overline{\nabla I}_t \overline{\nabla I}_t^T). \tag{2.29}$$

The variance σ_j is computed using the maximum likelihood.

2.5 Results

2.5.1 Crowded Scenes

Fig. 2.6 shows frames from eight videos of crowded scenes that we use for different applications in this dissertation. Videos of the concourse 2.6(a) and sidewalk 2.6(c) scenes are courtesy of Nippon Telegraph and Telephone Corporation. The faces in the these two scenes have been blurred to conceal identities. The sidewalk 2.6(c) and intersection 2.6(d) scenes are from the UCF Crowd Dataset [1]. The platform 2.6(e) and escalator 2.6(f) scenes are from Radke et al. [19]. The street 2.6(g) scene is from Mehran et al. [56], who also provide the synthetic video of a crowd interacting shown

(a) UMN1 (b) UMN2 (c) UMN3

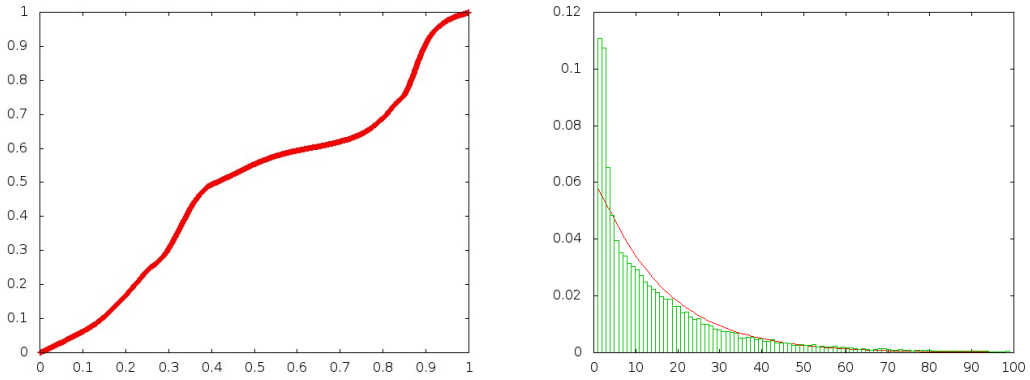
(d) UCSD1 (e) UCSD2

Figure 2.7: Frames from the UMN crowd dataset [73] (top) and the UCSD crowd dataset [72] (bottom) that have low densities.

in 2.6(h). Except the two scenes from the Nippon Telegraph and Telephone Corporation, the videos are available from the respective authors.

The density of the crowds, as well as the relative size of pedestrians, varies among the different scenes. The concourse and ticket gate scenes contain high density crowds captured at a near view. The sidewalk and street scenes are also captured at a close view, but are less dense. All four of these scenes contain pedestrians moving in an unstructured environment, resulting in a large variation of possible local motion patterns. The intersection, platform, escalator, and synthetic scenes all contain inherent structure: pedestrians are moving in specific directions due to the environment. Such scenes still contain variability due to the individuality of pedestrians.

Fig. 2.7 shows frames from five more scenes of crowds that have a low density. The first three (2.7(a)-(c)) are from the University of Minnesota (UMN) crowd dataset [73]. For each scene, the authors provide a number of clips of "normal" crowd behavior where pedestrians walk around, and



(a) (b)

Figure 2.8: The longitude plot (a) of $\boldsymbol{\mu}$ is near linear, suggesting that $\boldsymbol{\mu}$ follows a von Mises-Fisher distribution. The histogram (b) for κ resembles a gamma distribution.

then artificial instances of crowd panic where pedestrians run out of the scenes. The other two scenes (2.7(d) and 2.7(e)) are from the University of California San Diego (UCSD) dataset [72]. These real-world scenes have low density crowds, but the authors provide ground truth for unusual events. The low density in such scenes make it difficult to learn a structured pattern, but we are still able to achieve significant results as we later demonstrate.

2.5.2 HyperParameter Experiments

Here, we investigate our assumption regarding the priors we apply to the local motion pattern parameters κ and $\boldsymbol{\mu}$. We use a clip from the concourse scene (containing over $200,000$ cuboids) to investigate the form of these distributions. Mardia and El-Atoum [53] show that the conjugate prior on the mean direction $\boldsymbol{\mu}$ is also a von Mises-Fisher distribution. Fig 2.8 shows the longitudinal plot [54] of $\boldsymbol{\mu}$ for all of the cuboids. The plot is near linear, suggesting that the vectors are symmetrically distributed about the mean direction and do indeed follow a von Mises-Fisher distribution [54].

Fig. 2.8(b) shows a histogram of κ from the example video, which resembles a Gamma distribution (plotted in red), but is a poor fit for small values. This is acceptable, however, since extremely small values of κ (such as those less than 5) indicate wide distributions (i.e., unknown motion) and may

be considered equivalent. Bangert et al. [7] present a conjugate prior for κ, but it does not reflect the histogram in Fig. 2.8(b).

Next, we investigate the independence of κ and $\boldsymbol{\mu}$ independent. Previous work [7] also assume that κ and $\boldsymbol{\mu}$ are independent, but do not provide evidence. To achieve this, we estimate histograms of κ and $\boldsymbol{\mu}$ and compute the mutual information. We discretize $\boldsymbol{\mu}$ using a geodesic sphere with 162 faces. The mutual information is 0.08, which is reasonably low suggesting we may model the parameters independently.

2.6 Summary

In this chapter, we have presented a model of the crowd motion that can be learned from a video of a crowded scene. By basing the model on local motion patterns in the form of directional distributions, the model is indifferent to the *number* of pedestrians within the scene. Instead, the complexity of the model depends on the modality of the different flows that occur at each spatial location, an inherent characteristic of the heterogeneous motion of the crowd. In addition, by training a collection of hidden Markov models, our method captures the crowd dynamics that vary in space and time. Next, we explore how to leverage this model for specific video analysis tasks on crowded scenes. We will show that in each application the space-time model of local motion patterns provides a drastic improvement over other methods.

Chapter 3: Unusual Event Detection

3.1 Local Deviations from the Crowd

In this chapter, we demonstrate that the crowd motion can represent the *steady-state* (i.e., the typical motions of pedestrians) within the crowd. Specifically, we leverage the scene-wide model of the crowd motion to detect unusual events in local areas of the scene.

Unusual event detection is a key application in automatic surveillance systems. The sheer number of surveillance cameras deployed produces an abundance of video that is often only viewed after an incident occurs. By automatically detecting disturbances within the scene, the automatic surveillance system can alert security personnel as soon as an incident occurs. While large-scale unusual events, such as stampedes, incidents of violence, and crowd panic, are a primary motivation for automatic video surveillance, they are not the only disturbance that may need to be detected in crowded scenes. Detecting local unusual events (consisting of only a few of the scene's constituents) is also a challenging problem, especially in very crowded scenes since they can easily go unnoticed or disguised due to the heavy clutter within the scene.

We detect local unusual events by training our model on a video of typical crowd activity (thereby encoding the steady-state of the crowd), and then identifying local motion patterns from a different video of the same scene that statistically deviate from the learned model. Our model allows us to detect unusual events that occur on the cuboid level, consisting of one or multiple pedestrians. Since we train our model on a video of typical crowd behavior, the detected unusual events that are specific to the scene that is being analyzed. In addition, our space-time model enables the detection of unusual events that are specific to the location within the video frame. We demonstrate that our method detects subtle yet important events in high-density crowds, such as motion against the usual flow of traffic or a lack of motion in otherwise high motion areas.

3.2 Comparison with Related Work

Many approaches to detecting unusual events model each undesirable event explicitly [16, 24], but this requires all the possible unusual events to be known a priori. The large amount of variability in crowded scenes makes specifically modeling each event extremely difficult, if not impossible. In addition, such an approach scales poorly as each event increases the computation time required for detection. In contrast, we use our crowd model to represent the usual activity within the scene, and detect unusual events as local motion patterns that deviate from the learned model.

Ali and Shah [1] segment the crowd into spatial areas with homogeneous motion. New segments that occur from pedestrians moving in different directions are identified as unusual events. Heterogeneous crowds or those captured at a near-view, however, exhibit many different motions that naturally evolve over time. In other words, the changes in the motion of pedestrians detected by Ali and Shah [1] can well be part of the usual state of the crowd.

Mehran et al. [56] measure the "social force," i.e., the influence of the crowd, on the pedestrian and then detect anomalous frames using Latent Dirichlet Allocation. Their measure of social force is the difference between the instantaneous optical flow and the optical flow averaged over a fixed window of time. Since they average the optical flow over the video clip, long unusual events (i.e., those that span the entire clip) are considered typical activity. In contrast, we train our models on a video of typical behavior, and use them to detect unusual events in videos of the same scene recorded at a different time.

Mahadevan et al. [51] use mixtures of dynamic textures to identify anomalies in surveillance videos. Using dynamic textures, however, retains appearance variations which can introduce noise into the model and degrade results. In addition, the dynamic textures are computed over the entire frame, requiring a saliency measure to identify where in the frame the anomaly occurred. We compare our results with that of Mahadevan et al. [51] in Sec. 3.4.2.

3.3 Local Unusual Events

The collection of HMMs represent the underlying steady-state motion of the crowd by the spatial and temporal variations of local motion patterns. We seek to identify if a specific local motion pattern from a different video of the same scene contains unusual pedestrian activity. For the purpose of this work, we consider a local motion pattern unusual if it occurs infrequently or is absent from the training video. We derive a probability measure of how much a specific local motion pattern deviates from the crowd motion in order to identify unusual events.

Deviations from the HMM are caused by either an unlikely transition between sequential local motion patterns or a local motion pattern with low emission probabilities. We identify unusual local motion patterns by thresholding the conditional likelihood

$$\mathcal{T}_t = \mathrm{p}\left(O_t | O_1, \ldots, O_T\right) \tag{3.1}$$

where T is the last local motion pattern in the video clip. Exploiting the statistical independence properties of the HMM yields

$$\mathcal{T}_t = \sum_j^J \mathrm{p}(O_t | s_t = j) \boldsymbol{\gamma}_t(j) \ , \tag{3.2}$$

where $\boldsymbol{\gamma}_t$ is the posterior $\mathrm{p}(s_t = j | O_1, \ldots, O_T)$ computed from the Forwards-Backwards algorithm (see Eq. 2.21). The posterior $\boldsymbol{\gamma}_t$ encodes the expected latent state given the temporal statistics of the crowd, and decreases $\mathcal{T}_t$ when there is an unlikely transition. The emission likelihoods $\mathrm{p}(O_t | s_t = j)$ are low for all values of j when O_t was absent from the training data.

Computing $\mathcal{T}_t$ requires the entire video clip to be available. In a real-world system, unusual events need to be detected as they occur. To achieve this, we use the predictive distribution to compute an alternative measure

$$\widetilde{\mathcal{T}}_t = \sum_j^J \mathrm{p}(O_t | s_t = j) \hat{\boldsymbol{\alpha}}_t(j) \tag{3.3}$$

where $\hat{\boldsymbol{\alpha}}_t$ is the posterior $\mathrm{p}(s_t = j | O_1, \ldots, O_t)$ computed during the forwards phase of the Forwards-

(a) (b)

Figure 3.1: Detection results of pedestrians moving against the crowd.

Backwards algorithm (see Eq. 2.20). The estimate $\widetilde{\mathcal{T}}_t$ only requires the local motion patterns up to time t to be available, but does not consider the transition *out* of the observation at time t. As such, some cuboids are incorrectly classified but anomaly detection can be performed online. We compare the results of both measures in Sec. 3.4.

Often, the crowd can display different modalities at different spatial locations of the scene. For example, some areas may regularly contain no motion, while others contain motion in multiple directions. As a result, the ideal threshold value may change with the spatial location. We account for this by dividing our likelihood measure $\widetilde{\mathcal{T}}_t$ by the average likelihood of the training data.

3.4 Results

After training these models on videos of consistent crowd motion, we detect unusual movements of pedestrian in query videos of the same scene recorded at a different time. For this chapter, we detect anomalies in the concourse and ticket gate scenes. The length of training videos varied for each example between 540 and $3,000$ frames, depending on the specific example. We use cuboids of size $30 \times 30 \times 20$ for the ticket gate scene and $40 \times 40 \times 20$ for the concourse scene.

(a) (b)

Figure 3.2: Detection of no motion in otherwise high motion areas.

3.4.1 Detected Unusual Events

Fig. 3.1 shows successful detection of unusual movements of pedestrians in local areas. True positives (i.e., detected unusual events) are shown in blue, false positives in pink, and false negatives in red. Fig. 3.1(a) shows three successful detections of pedestrians in the concourse scene moving from left to right against the regular crowd traffic. The training video used for the examples consists of pedestrians moving in many different directions, but not from the left side of the scene to the right. Fig. 3.1(b), from the ticket gate scene, shows three detections of pedestrians reversing directions in the turnstiles. These examples illustrate the unique ability of our approach to detect irregular local motion patterns within a crowded scene comprised of diverse movements of pedestrians. In addition, these examples illustrate that our approach is scene-specific: the unusual events that are detected in the ticket gate scene (i.e., moving from the bottom to the top of the frame) are usual activity in the concourse scene.

The type of detected events depends entirely on the training data. Fig. 3.2 shows detection of areas lacking movement in otherwise high-motion areas. Since the training video contains typical crowd motion, the lack of pedestrians (e.g., the empty turnstiles in Fig. 3.2(b) or visible background in Fig. 3.2(a)) deviate from the model. This dependency on the training data is not only expected, it is desirable. It allows users of our approach to decide which particular local movements of pedestrian they consider usual by including it in the training video. Note also that the detection is *location specific*: the areas that typically lack motion (upper corners in both frames) are correctly classified as usual.

Fig. 3.3 shows the receiver operating characteristic (ROC) curves (generated by varying the

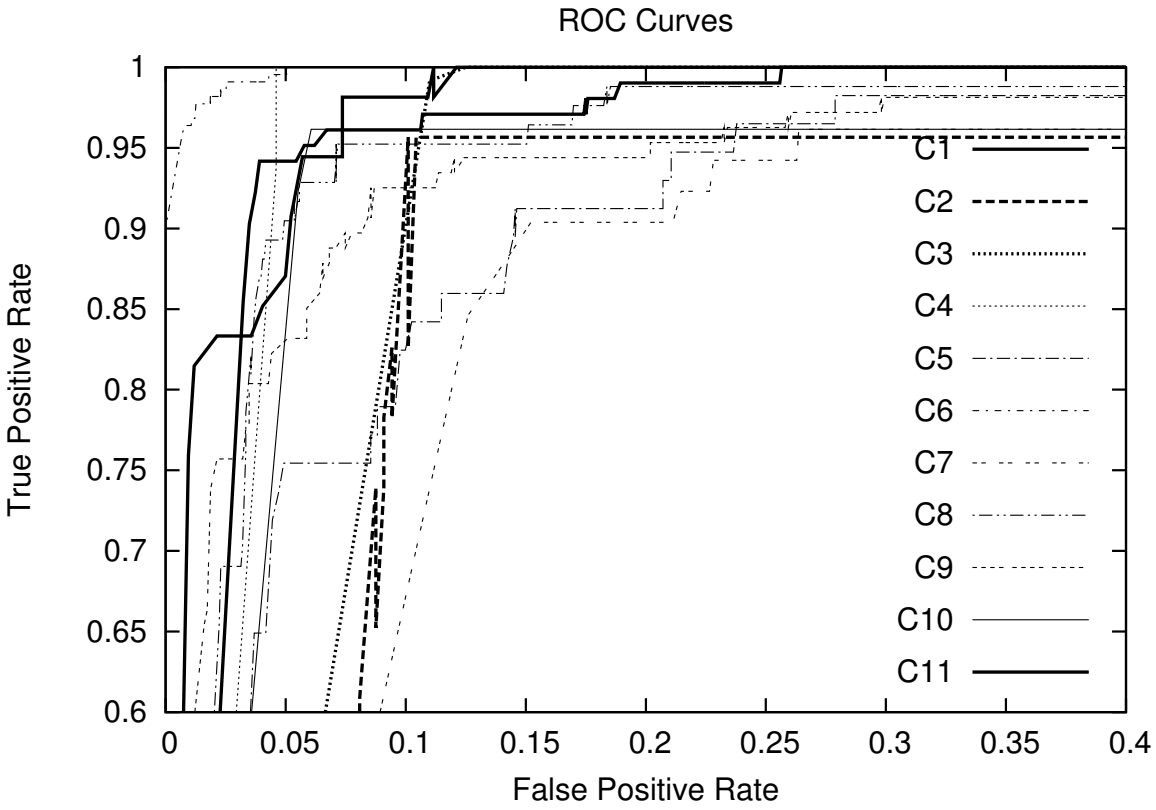


Figure 3.3: Anomaly detection receiver operating characteristic curves for 11 clips.

likelihood threshold) for all of the clips. Our approach performs with significant accuracy on each of the example videos. The clip C7 contains multiple simultaneous unusual events, which are more difficult to detect accurately.

Fig. 3.4 shows the detection accuracy (the average of the true positive rate and the true negative rate) using the online likelihood measure (computed from $\boldsymbol{\alpha}$) compared with the full likelihood measure (computed from $\boldsymbol{\gamma}$). The online method performs comparably to the offline method in all but one clip (C4). This clip contains a single pedestrian moving against the crowd, and is shown in the bottom left of Fig. 3.1(a). In this case, the extra temporal information yields more accurate results due to the subtle nature of the unusual event.

3.4.2 Comparison to Other Methods

It is important to distinguish crowded scenes that lack temporal variations, such as those shown in Fig. 3.5 from Ali and Shah [1], from those with significant variations in local motion patterns. The

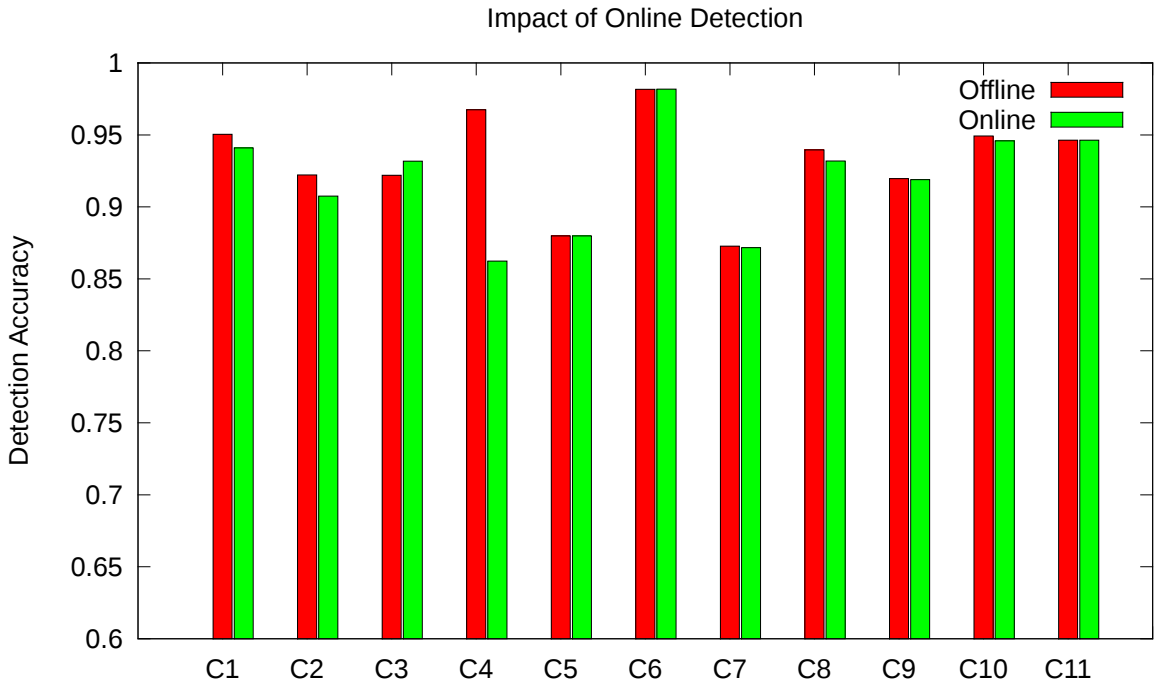
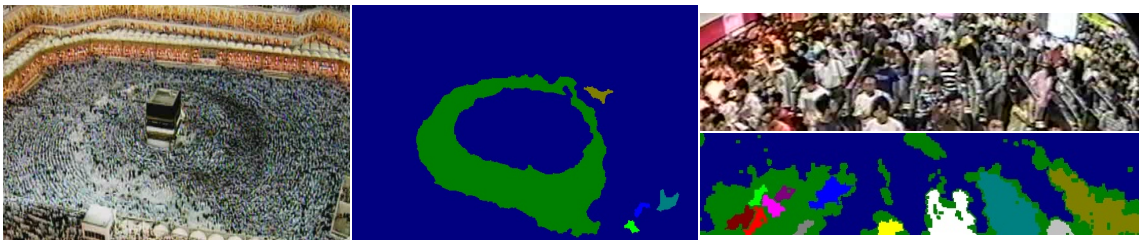


Figure 3.4: Detection accuracy using online and offline method for 11 clips.

movement of pedestrians in the concourse and ticket gate scenes varies dramatically over different areas of the frame and throughout the video. To illustrate this difference, Fig. 3.5(c) shows the crowd flow segmentation of Ali and Shah [1] on an 80 frame clip of the ticket gate scene. The different movements of pedestrians produce diverse flow vectors that drastically affect the segmentation. Areas with a large amount of variations, such as the lower left of the scene, are over-segmented and do not reflect the motion of the crowd. Distant areas of the scene are in under-segmented, disregarding the diverse possible motions of the pedestrians before they enter the ticket gate.

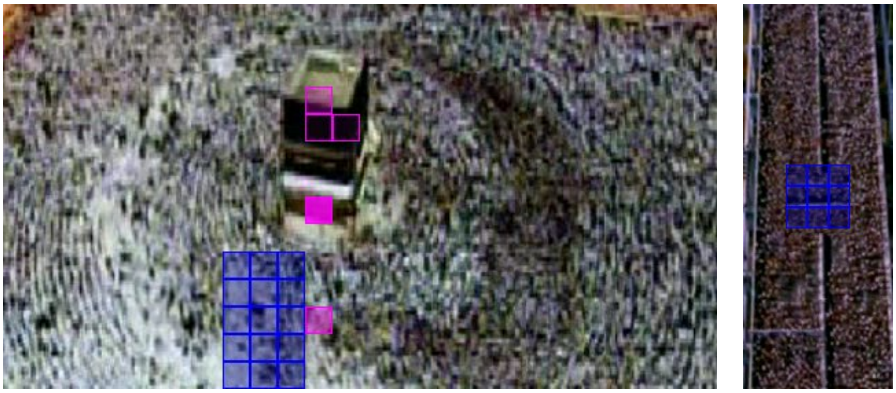
Fig. 3.6 shows our detection result of "unstable" crowd activity in the videos used by Ali and Shah [1]. The local motion patterns in these scenes lack the temporal variations present in dynamically varying crowded scenes. In addition, the unusual events in these videos are artificial modifications of the training video. As a result, our method has high detection accuracy.

Fig. 3.7 shows the detection accuracy on 9 clips from the UCSD dataset [72]. We use the results of Mahadevan et al. [51] posted on the web [52] for comparison. Our approach achieves a higher detection accuracy for all of the clips. Fig. 3.8 shows detection results using our method and that of



(a) (b) (c)

Figure 3.5: Segmentation of crowded scenes using the approach of Ali and Shah [1].



(a) (b)

Figure 3.6: Detection of unstable crowd activity in videos from Ali and Shah [1].

Mahadevan et al. [51]. The dynamic textures and saliency measure used by Mahadevan et al. [51] is unable to localize the unusual event with high accuracy. In contrast, our method classifies each individual cuboid, resulting in a higher number of correctly classified cuboids.

Fig. 3.9 shows the effects of increasing the training data size for video C1. As expected, the performance increases with longer training data, and achieves good performance with 100 observations, or $2,000$ frames of video. Using only 50 observations the model achieves significant accuracy with a false positive rate (ratio of false positives to total negatives) of 0.17 and a true positive rate (ratio of true positives to total positives) of 0.88. This strong performance with few observations directly results from the crowd's high density. Since the scene contains a large number of pedestrians, significant variations in local motion patterns occur even in short video clips.

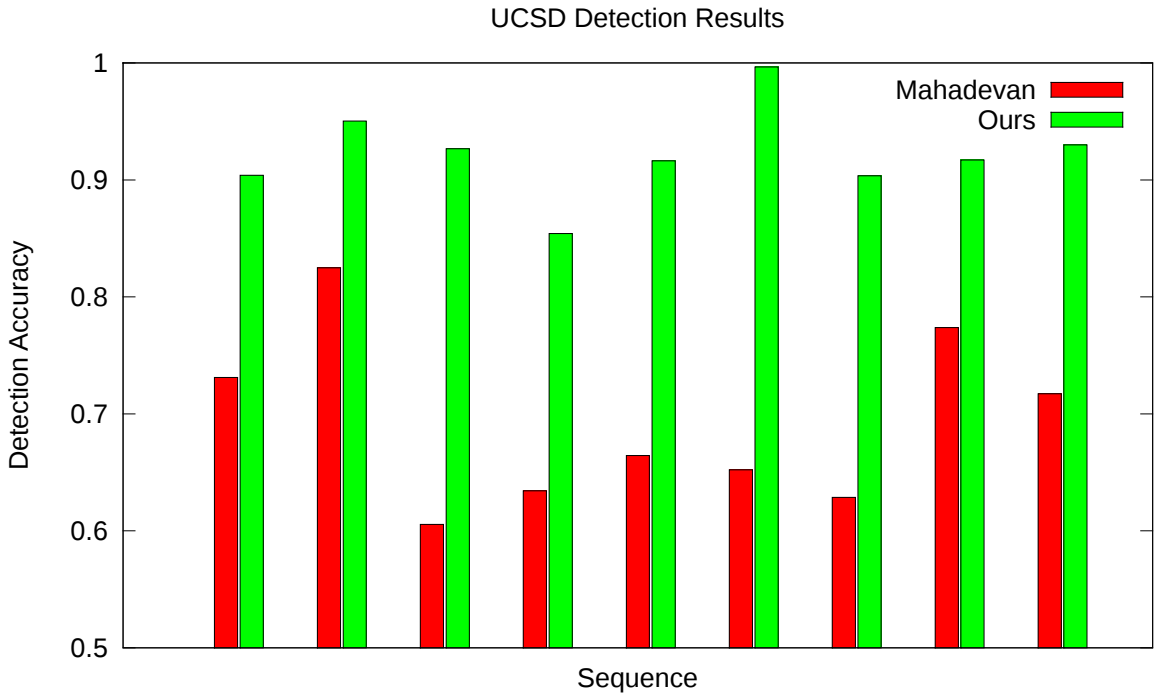


Figure 3.7: Detection accuracy on the UCSD dataset using our method compared with that of Mahadevan et al. [51].

3.5 Summary

In this chapter, we demonstrated how our model may be used to represent the steady-state, i.e., the typical spatio-temporal variations in motion, of the crowd. To this end, we presented a method for identifying local unusual events as statistical deviations. Since our model can be learned from an example video, our method detects anomalies that are specific to the scene being analyzed. In addition, the detected unusual events are location specific: they depend on where in the scene the anomaly occurs. Next, we explore how the crowd motion model may be leveraged as an indicator of an individual's motion.

(a) (b)

Figure 3.8: Example detection results using our method (a) and that of Mahadevan et al. [51] (b).

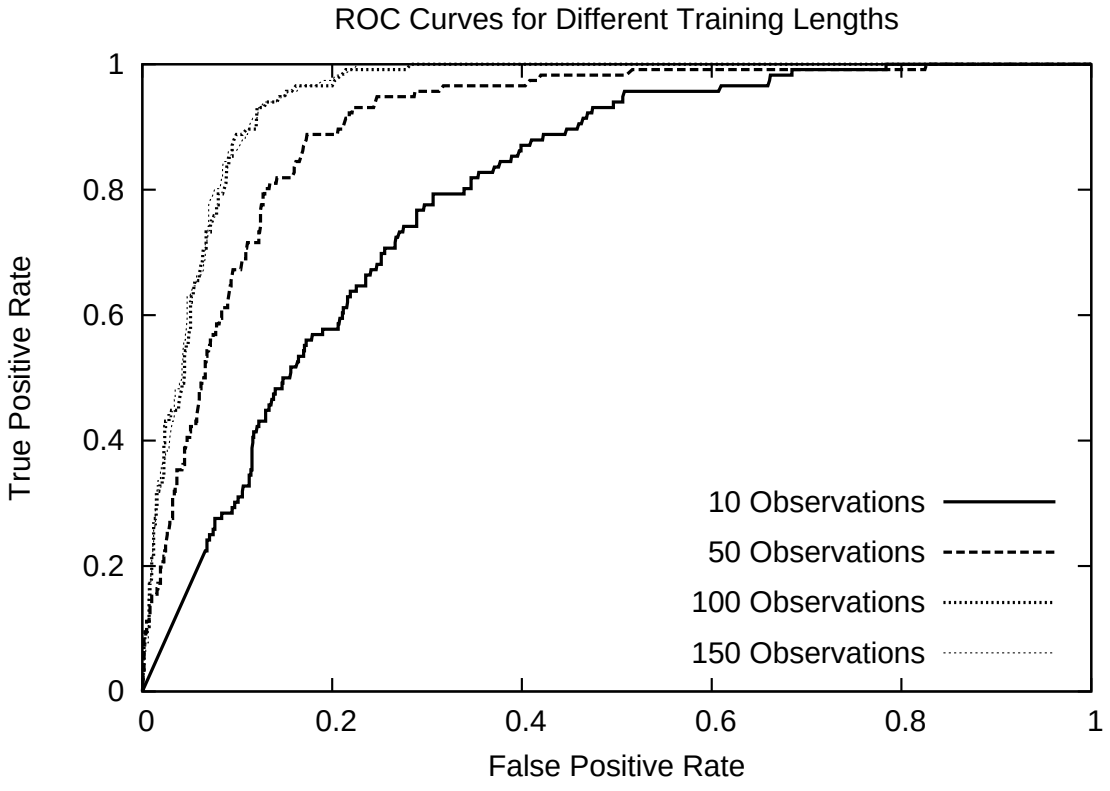


Figure 3.9: Effects of increasing training data.

Chapter 4: Tracking Individual Pedestrians

4.1 Using the Crowd to Track Individuals

In this chapter, we demonstrate how the crowd motion can be leveraged to indicate the movement of a single pedestrian. Specifically, we use the scene-wide model of the crowd as a prior to track individuals.

Tracking objects or people is a crucial step in video analysis with a wide range of applications including behavior modeling and surveillance. The complex motions of pedestrians and lack of visible background make tracking in crowded scenes difficult. To overcome these challenges, we use our model of the crowd motion to predict the motion of individuals moving through videos of crowded scenes. We use the crowd motion to predict local motion patterns in videos containing the pedestrian that we wish to track. Next, we use the predicted local motion pattern as a prior on the state-transition distribution in a particle filter framework. We show that our approach accurately predicts the motion that a target will exhibit during tracking. By use our space-time model, we achieve more accurate tracking results compared with previous methods that disregard temporal information.

4.2 Comparison with Related Work

Previous tracking work are mostly object-centric, based on the modeling of the motion and appearance of the target individual. Recent work extend these methods to handle crowded scenes by deriving a robust appearance model of each individual. A representative approach by Zhao et al. [81] tracks multiple pedestrians in videos of crowds by modeling the camera, human shapes, and the background's appearance. Though improved accuracy can be achieved by a more descriptive model of each tracked target, the large amount of partial occlusions and significant variations in the movement of pedestrians in videos of dense crowds still present significant challenges to such approaches.

Techniques that track multiple objects, such as that of Hue et al. [35], detect interest points within each frame to describe the objects within the scene. Tracking is performed concurrently on all of the objects by establishing correspondences between the points in different frames. As noted by Khan et al. [40], a single point may be shared between multiple targets and present ambiguity to the tracker. Shared points are often the result of motion boundaries or clutter, both of which occur frequently in videos of crowded scenes. Rather than associating points between frames, our approach predicts the motion a single target will exhibit between frames. Khan et al. [39] model the interaction among the interest points to improve the tracking of each object. Due to the large number of pedestrians within crowded scenes, modeling the interaction among and the separate motion of individual objects is intractable. By using a scene-centric model of the crowd motion we can track an individual without explicitly modeling each moving object.

Ali and Shah [2] track pedestrians in videos of crowded scenes captured at a distance using a number of "floor fields" that describe how a pedestrian should move based on scene-wide constraints. Their sink-seeking and boundary floor fields, which describe the exit points and motion-boundary areas of the frame, respectively, impose a single likely direction at each spatial location of the video. In contrast, we model the crowd motion as a temporally evolving system that allows for any number of likely movements in each space-time location of the video. Though their dynamic floor field allows for some variation with the moving pedestrian, it is computed using frames after the current tracking position. Our approach uses the previously observed frames of the video to predict the motion of the target, allowing it to operate online. We compare the results of our method to that of Ali and Shah in Section 4.6.

Rodriguez et al. [62] use a topical model to represent motion in different directions at each spatial location. This approach imposes a fixed number of motion directions at each spatial location, but disregards the temporal relationship between sequentially occurring local motions that videos of crowded scenes naturally exhibit. In addition, Rodriguez et al. quantize the optical flow vectors into 10 possible directions, while our approach uses a full distribution of optical flow for a more robust and descriptive representation of the motion of tracked targets. Such a coarse quantization

limits tracking in crowded scenes to only a few directions. A finer quantization results in diminishing returns (as discussed by Rodriguez et al. [62]). We directly compare the results of our method to that of Rodriguez et al. [62] in Section 4.6.

4.3 Predicting Motion Patterns

We train our model of the crowd motion on a video of a crowded scene containing typical crowd behavior. Next, we use it to predict the local motion patterns at each location of a different video of the same scene. Note that, since we create a scene-centric model based on the changing motion in local regions, the prediction is independent of which individual is being tracked. In fact, we predict the local motion pattern at all locations of video volume given only the previous frames of the video.

Given a trained HMM at a specific spatial location and a sequence of observed local motion patterns $O_1, \ldots, O_{t-1}$ from the query video, we seek to predict the next local motion pattern $\tilde{O}_t = \{\tilde{\boldsymbol{\mu}}_t, \tilde{\kappa}_t\}$ that will occur. We achieve this by computing the expected value of the predictive distribution, i.e.,

$$\tilde{O}_t = \mathbb{E}\left[\mathrm{p}\left(O_t | O_1, \ldots, O_{t-1}\right)\right] \ . \tag{4.1}$$

Via marginalization

$$\mathrm{p}\left(O_t | O_1, \ldots, O_{t-1}\right) = \sum_j^J \mathrm{p}\left(O_t | s_t = j\right) \sum_i^J \mathrm{p}\left(s_t = j | s_{t-1} = i\right) \mathrm{p}\left(s_{t-1} = i | O_1, \ldots, O_{t-1}\right) \ . \tag{4.2}$$

Note that $\mathrm{p}\left(s_{t-1} = i | O_1, \ldots, O_{t-1}\right)$ is the posterior $\hat{\boldsymbol{\alpha}}_{t-1}$ computed during the Forwards-Backwards algorithm [9] (see Eq. 2.20), and $\mathrm{p}\left(s_t = j | s_{t-1} = i\right)$ is defined by the HMM's state transition matrix $\mathbf{A}$. As such, the second summation in Eq. 4.2 may be represented by

$$\boldsymbol{\omega}_t(j) = \sum_i^J \mathbf{A}(i,j) \hat{\boldsymbol{\alpha}}_{t-1}(i) \tag{4.3}$$

and

$$\tilde{O}_t = \sum_j^J \mathbb{E}\left[\mathrm{p}\left(O_t | s_t = j\right)\right] \boldsymbol{\omega}_t(j) \ . \tag{4.4}$$

Thus $\tilde{O}_t$ is a weighted sum of the expected local motion patterns defined by each emission density.

As discussed in Sec. 2.4.2, each emission density $\mathrm{p}\left(O_t|s_t=j\right)$ is defined by four parameters $\left\{\boldsymbol{\mu}_0^j, \kappa_0^j, a^j, \theta^j\right\}$. Using the mean of the Gamma and von Mises-Fisher distributions

$$\mathbb{E}\left[\mathrm{p}\left(O_t|s_t=j\right)\right] = \left\{\boldsymbol{\mu}_0^j, a^j\theta^j\right\} \ , \tag{4.5}$$

i.e., a local motion pattern with mean direction $\boldsymbol{\mu}_0^j$ and concentration parameter $a^j\theta^j$. Thus the predicted local motion pattern $\tilde{O}_t$ is defined by mean direction

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\left|\sum_j^J \boldsymbol{\omega}_t(j)\boldsymbol{\mu}_0^j\right|} \sum_j^J \boldsymbol{\omega}_t(j)\boldsymbol{\mu}_0^j \tag{4.6}$$

and concentration parameter

$$\tilde{\kappa}_t = \sum_j^J \boldsymbol{\omega}_t(j)a^j\theta^j \ . \tag{4.7}$$

During tracking, we use the previous frames of the video to predict the local motion pattern that spans the next M frames (where M is the number of frames in a cuboid). Since the predictive distribution is a function of the HMM's transition probabilities and the hidden states' posteriors, the prediction may be computed on-line and efficiently during the forward phase of the Forwards-Backwards algorithm [60].

4.4 Bayesian Tracking

We now use the predicted local motion pattern to track individuals in a Bayesian framework. Specifically, we use the predicted local motion pattern as a prior on the parameters of a particle filter. Our model of the crowd motion, i.e., the collection of HMMs, enables these priors to vary in the space-time and dynamically adapt to the changing motions within the crowd.

Tracking can be formulated in a Bayesian framework [36] by maximizing the posterior distribution of the state $\mathbf{x}_f$ of the target at frame f given past and current measurements $\mathbf{z}_{1:f} = \{\mathbf{z}_i|i = 1 \ldots f\}$. Note that the index of each frame f is different from the temporal index t of the local motion

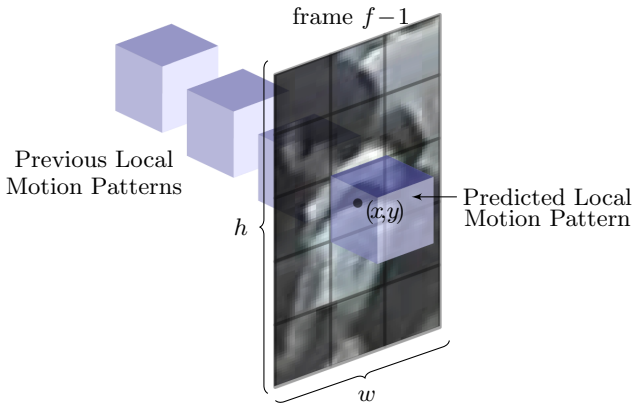


Figure 4.1: We use the predicted local motion pattern to impose a prior on the motion of the pedestrian through the space-time volume.

patterns (since the cuboids span many frames). We define state $\mathbf{x}_f$ as a four-dimensional vector $[x, y, w, h]^T$ containing the tracked target's 2D location (in image space), width, and height, respectively. Tracking is performed by maximizing the posterior distribution

$$\mathrm{p}(\mathbf{x}_f|\mathbf{z}_{1:f}) \propto \mathrm{p}(\mathbf{z}_f|\mathbf{x}_f) \int \mathrm{p}(\mathbf{x}_f|\mathbf{x}_{f-1})\, \mathrm{p}(\mathbf{x}_{f-1}|\mathbf{z}_{1:f-1})\, d\mathbf{x}_{f-1}\,, \tag{4.8}$$

where $\mathbf{z}_f$ is the frame at time f, $p(\mathbf{x}_f|\mathbf{x}_{f-1})$ is the transition distribution, $p(\mathbf{z}_f|\mathbf{x}_f)$ is the likelihood, and $p(\mathbf{x}_{f-1}|\mathbf{z}_{1:f-1})$ is the posterior from the previous tracked frame. The transition distribution $p(\mathbf{x}_f|\mathbf{x}_{f-1})$ models the motion of the target between frames $f-1$ and f, and the likelihood distribution $p(\mathbf{z}_f|\mathbf{x}_f)$ represents how well the observed image $\mathbf{z}_f$ matches the state $\mathbf{x}_f$. Often, the distributions are non-Gaussian, and the posterior distribution is estimated using a Markov chain Monte Carlo method such as a particle filter [36] (please refer to [5] for an introduction to particle filters).

As shown in Fig. 4.1, we impose priors on the transition $p(\mathbf{x}_f|\mathbf{x}_{f-1})$ distribution using the predicted local motion pattern at the space-time location defined by $\mathbf{x}_{f-1}$. For computational efficiency, we use the cuboid at the center of the tracked target to define the priors, although the target may span several cuboids across the frame.

4.4.1 Transition Distribution

We use the predicted local motion pattern to hypothesize the motion of the tracked target between frames $f-1$ and f, i.e., the transition distribution $p(\mathbf{x}_f|\mathbf{x}_{f-1})$. Let the state vector $\mathbf{x}_f = \left[\mathbf{k}_f^T, \mathbf{d}_f^T\right]^T$ where $\mathbf{k}_f = [x, y]$ is the target's location (in image coordinates) and $\mathbf{d}_f = [w, h]$ is the size (width and height) of a bounding box around the target. We focus on the target's movement between frames and use a second-degree auto-regressive model [59] for the transition distribution of the size $\mathbf{d}_f$ of the bounding box.

The transition distribution of the target's location $\mathrm{p}(\mathbf{k}_f|\mathbf{k}_{f-1})$ reflects the 2D motion of the target between frames $f-1$ and f. We model this using the von Mises-Fisher distribution defined by the predicted local motion pattern $\tilde{O}_t = \{\tilde{\boldsymbol{\mu}}_t, \tilde{\kappa}_t\}$ at space-time location $\mathbf{k}_{f-1}$. In the particle filter, a set of N sample locations (i.e., particles) $\{\mathbf{k}_{f-1}^i | i = 1, \ldots, N\}$ are drawn from the prior $\mathrm{p}(\mathbf{x}_{f-1}|\mathbf{z}_{1:f-1})$. For each sample $\mathbf{k}_{f-1}^i$, we draw a 3D flow vector $\mathbf{v}^i = [v_x^i, v_t^i, v_t^i]$ from the predicted local motion pattern at space-time location $\mathbf{k}_{f-1}^i$. We use these 3D flow vectors to update each particle

$$\mathbf{k}_f^i = \mathbf{k}_{f-1}^i + \begin{bmatrix} v_x^i / v_t^i \\ v_y^i / v_t^i \end{bmatrix}. \tag{4.9}$$

Note that $\tilde{\kappa}_t$ plays a key role in this step: distributions with a large variance will spread the particles over the frame, while those with a small variance (i.e., determinable flow) will keep the particles close together.

In summary, we use the predicted local motion pattern as the transition distribution at the space-time location of the target. By doing so, the state-transition distribution is robust to unreliable flow estimates and dynamically adapts to the space-time dynamics of the crowd.

4.4.2 Likelihood Distribution

Typical models of the likelihood distribution maintain a template T that represents the target's characteristic appearance in the form of a color histogram [59] or an image [2]. A template T and the region R (the bounding box defined by state $\mathbf{x}_f$) of the observed image $\mathbf{z}_f$ are used to model

the likelihood distribution

$$\mathrm{p}\left(\mathbf{z}_f|\mathbf{x}_f\right) = \frac{1}{Z}\exp\left[\frac{-d\left(R,T\right)^2}{2\sigma^2}\right], \tag{4.10}$$

where σ is the variance selected empirically, $d(\cdot)$ is a distance measure, and Z is a normalization constant.

Rather than using color histograms [59] or intensity [2] as the defining characteristic of an individual's appearance, we model the template T as an image of the individual's spatio-temporal gradients. This representation is more robust to appearance variations caused by noise or illumination changes. We use a weighted sum of the angles between the spatio-temporal gradient vectors in the observed region and the template to define the distance measure

$$d\left(R,T\right) = \sum_{i}^{M} \rho_i^f \arccos\left(\mathbf{t}_i^T \mathbf{r}_i\right), \tag{4.11}$$

where M is the number of pixels in the template, $\mathbf{t}_i$ is the normalized spatio-temporal gradient vector in the template, $\mathbf{r}_i$ is the normalized spatio-temporal gradient vector in the region R of the observed image at frame f, and ρ_i^f is the weight of the pixel at location i and frame f,

We model changes in the target's appearance by estimating the weights $\left\{\rho_i^f|i=1,\ldots,M\right\}$ in Eq. 4.11 during tracking. Specifically, pixels that change drastically (due to the pedestrian's body movement or partial occlusions) exhibit a large error between the template and the observed region. We estimate this error E_i^f during tracking to account for a pedestrian's changing appearance. The error at frame f and pixel i is

$$E_i^f = \alpha \arccos\left(\mathbf{t}_i^T \mathbf{r}_i\right) + \left(1-\alpha\right) E_i^{f-1}, \tag{4.12}$$

where α is the update rate (set to 0.05) and $\mathbf{t}_i$ and $\mathbf{r}_i$ are again the gradients of the template and observed region, respectively. To reduce the contributions of frequently changing pixels to the

distance measure, we use the error at frame f and pixel i to compute the weight

$$\rho_i^f = \frac{1}{Z}\left(\pi - E_i^{f-1}\right), \tag{4.13}$$

where Z is a normalization constant such that $\sum_i \rho_i^f = 1$.

To account for changes in appearance, the template is updated each frame by a weighted average also using $\alpha = 0.05$.

4.5 Gaussian Local Motion Patterns

As discussed in Sec. 2.3, we use directional distributions to represent the local motion patterns but our previously published tracking work [43] used 3D Gaussian distributions. Here, we show how to predict local motion patterns using the Gaussian representation, and how to use them to hypothesize the state transition distribution. In addition, since there is some appearance information in the 3D Gaussian representation, we can impose a prior on the likelihood distribution $\mathrm{p}\left(\mathbf{x}_f|\mathbf{z}_{1:f}\right)$ as well. We emphasize that, in practice, the directional distributions result in more accurate predictions with a lower dimensional model, but we detail the use of Gaussian distributions for completeness.

If the 3D Gaussian representation is used, the expected value of the emission probability $p\left(O_t|s_s = j\right)$ is the local motion pattern $P_j = \left\{\overline{\nabla I}_j, \widetilde{\mathbf{\Sigma}}_j\right\}$. Therefore, the predicted local motion pattern (defined by $\widetilde{\nabla I}_t$ and $\widetilde{\mathbf{\Sigma}}_t$) is a weighted sum of the set of 3D Gaussian distributions associated with the HMM's hidden states

$$\widetilde{\nabla I}_t = \sum_j^J \boldsymbol{\omega}_t(j)\overline{\nabla I}_j\,, \tag{4.14}$$

$$\widetilde{\mathbf{\Sigma}}_t = -\widetilde{\nabla I}_t\widetilde{\nabla I}_t^T + \sum_j^J \boldsymbol{\omega}_t(j)\left(\widetilde{\mathbf{\Sigma}}_j + \overline{\nabla I}_j\overline{\nabla I}_j^T\right), \tag{4.15}$$

where $\boldsymbol{\omega}_t$ is the prediction weight given in Eq. 4.3.

The state transition distribution is normal and defined by

$$p\left(\mathbf{k}_f|\mathbf{k}_{f-1}\right) = \mathcal{N}\left(\mathbf{k}_f - \mathbf{k}_{f-1}; \mathbf{u}, \mathbf{\Lambda}\right), \tag{4.16}$$

where $\mathbf{u}$ is the 2D optical flow vector and $\mathbf{\Lambda}$ is its covariance matrix. We estimate the parameters $\mathbf{u}$ and $\mathbf{\Lambda}$ from the predicted local motion pattern at the space-time location containing $\mathbf{k}_{f-1}$. First, we construct the predicted structure tensor

$$\widetilde{\mathbf{G}} = \widetilde{\mathbf{\Sigma}} + \widetilde{\nabla} I_t \widetilde{\nabla} I_t^T, \tag{4.17}$$

which we use to estimate the optical flow as in Eq. 2.3.

Next, we estimate the covariance matrix $\mathbf{\Lambda}$ of the 2D optical flow from the same predicted structure-tensor matrix $\widetilde{\mathbf{G}}$. The 3D optical flow $\mathbf{q}$ is orthogonal to the plane containing the space-time gradients. The ordered eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of $\widetilde{\mathbf{G}}$ encode a confidence of the optical flow estimate [76]. If the video volume contains a single dominant motion vector and sufficient texture, then $\widetilde{\mathbf{G}}$ is close to rank 2 and $\lambda_3 \ll \lambda_1, \lambda_2$. We model the confidence in the optical flow estimate as inversely proportional to its variance, and thus consider the eigenvalues of $\widetilde{\mathbf{G}}$ inversely proportional to the eigenvalues of $\mathbf{\Lambda}$. If $\widetilde{\mathbf{G}}$ is close to full rank, then the gradients are not co-planar and the 3D optical flow may vary in the primary directions of the gradient distributions (i.e., the other eigenvectors of $\widetilde{\mathbf{G}}$). The optical flow vector $\mathbf{u}$ is a projection of the 3D optical flow $\mathbf{q}$ onto the plane $f = 1$. We let the eigenvectors of $\mathbf{\Lambda}$ be the projection of the other two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\widetilde{\mathbf{G}}$ onto the same plane. This encodes uncertainty in the primary directions of the spatio-temporal gradients. Therefore, we estimate $\mathbf{\Lambda}$ by

$$\mathbf{\Lambda} = [\mathbf{v}_1', \mathbf{v}_2'] \begin{bmatrix} \frac{\lambda_3}{\lambda_1} & 0 \\ 0 & \frac{\lambda_3}{\lambda_2} \end{bmatrix} [\mathbf{v}_1', \mathbf{v}_2']^{-1}, \tag{4.18}$$

where $\mathbf{v}_1'$ and $\mathbf{v}_2'$ are the projections of $\mathbf{v}_1$ and $\mathbf{v}_2$ onto the plane $f = 1$. The 2D covariance matrix $\mathbf{\Lambda}$ implies a larger variance when the optical flow estimate is poor, and a smaller variance when the optical flow estimate is reliable.

Next, we use the predicted local motion pattern as a prior on the likelihood distribution $p(\mathbf{z}_f|\mathbf{x}_f)$. Specifically, we hypothesize the variance σ^2 (in Eq. 4.10) to account for further variations in the

pedestrian's appearance (i.e., due to the individual's natural body movement or partial occlusions). Intuitively, tracking a pedestrian whose appearance changes frequently requires a larger variance to account for more variation and avoid drift. Alternatively, tracking a pedestrian whose appearance changes infrequently benefits from a smaller variance for more accurate results. Similar to the transition distribution, we hypothesize σ^2 using the predicted local motion pattern at the space-time location that the state $\mathbf{x}_{f-1}$ occurs. The covariance of the predicted local motion pattern is a 3×3 matrix

$$\widetilde{\mathbf{\Sigma}}_t = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} \tag{4.19}$$

of spatio-temporal gradients. We consider σ_{zz} (i.e., the variance of the temporal gradient) representative of the amount of appearance change that occurs within the cuboid: temporal gradients from frequent appearance changes have a larger variance. The degree of change, however, depends on the appearance of the specific target. Intuitively, targets with a large amount of texture will naturally exhibit local motion patterns with larger values of σ_{zz}. Thus, we consider σ_{zz} proportional to the variance σ^2 of the likelihood distribution, and estimate it by

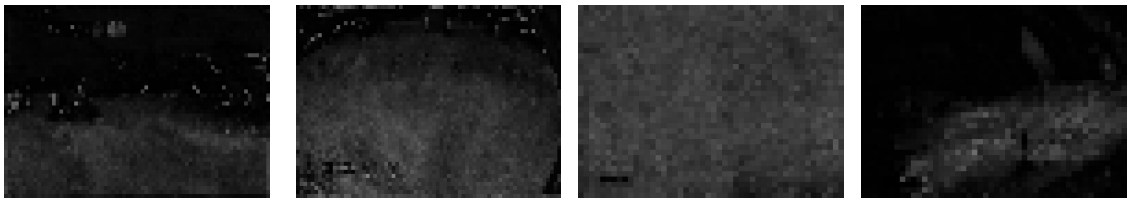
$$\sigma^2 = C\sigma_{zz}^2 \,, \tag{4.20}$$

where C is a scaling factor selected empirically (typically results are satisfactory letting $C = 50$).

To summarize, using the 3D Gaussian representation allows us to dynamically adjust the likelihood distribution to account for appearance variations. The excess appearance information stored in the 3D Gaussian representation, however, causes the predicted optical flow to be less accurate.

4.6 Results

We evaluate our method on videos of four scenes: the concourse and ticket gate scenes, and the sidewalk and intersection scenes from the UCF dataset [1] (see Sec. 2.5.1 for example frames). We use a sampling importance re-sampling particle filter as in [36] with 100 to 800 particles (depending



(a) Concourse (b) Ticket Gate (c) Sidewalk (d) Intersection

Figure 4.2: The angular error between the predicted optical flow vector and the observed optical flow vector for all scenes.

on the subject) to estimate the posterior in Eq. 4.8. We train a collection of HMMs on a video of each scene, and use it to track pedestrians in videos of the same scene recorded at a different time. The training videos for each scene have 300, 350, 300, and 120 frames, respectively. The training videos for the concourse, ticket gate, and sidewalk scenes have a large number of pedestrians moving in a variety of directions. The video for the intersection scene has fewer frames due to the limited length of video available. In addition, the training video of the intersection scene contains only a few motion samples in specific locations, as many of the pedestrians have moved to other areas of the scene in that point in time. Such sparse samples, however, still result in a useful model since most of the pedestrians are only moving in one of two directions (either from the lower left to the upper right, or from the upper right to the lower left).

Due to the perspective projection of many of the scenes, which is a common occurrence in surveillance, the sizes of pedestrians varies immensely. As such, the initial location and size of the targets are selected manually. Methods for automatically detecting pedestrians and their sizes [21, 47] may be used to automatically initialize our method.

The motion represented by the local motion pattern depends directly on the size of the cuboid. We use a cuboid of size $10 \times 10 \times 10$ on all scenes so that a majority of the cuboids are smaller than the space-time region occupied by a moving pedestrian. By doing so, the cuboids represent the motion of a single pedestrian but still contain enough pixels to accurately estimate a distribution of optical flow vectors.

We measure the accuracy of the predicted local motion patterns by the angle between the pre-

Figure 4.3: An illustration of the predicted optical flow changing over space and time.

Figure 4.4: An example of the changing predicted optical flow.

dicted flow $\tilde{\mu}_t$ and the observed optical flow. Fig. 4.2 shows the angular error averaged over the entire video for each spatial location in all four scenes. Dark areas indicate low error, and white areas indicate high error. The sidewalk scene contains errors in scattered locations due to the occasional visible background in the videos and close-view of pedestrians. Similarly, the closer locations in the ticket gate scene contain higher variability, and thus have a higher error.

Fig. 4.3 shows the predicted optical flow, colored by key in the lower left, for four frames from the sidewalk scene. Pedestrians moving from left to right are colored red, those moving right to left are colored green, and those moving from the bottom of the screen to the top are colored blue. As time progresses, our space-time model dynamically adapts to the changing motions of pedestrians within the scene as shown by the changing cuboid colors over the frames. Poor predictions appear as noise, and occur in areas of little texture such as the visible areas of the sidewalks or pedestrians with little texture.

Fig. 4.4 shows a specific example of the changing predicted optical flow on six frames from the sidewalk scene. In the first two frames the predicted flow is from the left to the right, correctly

corresponding to the motion of the pedestrian. In later frames the flow adjusts to the motion of the pedestrian at that point in time. Only by exploiting the temporal structure within the crowd motion are such dynamic predictions possible.

Fig. 4.5 shows a visualization of our tracking results on videos from each of the different scenes. Each row shows 4 frames of our method tracking different targets whose trajectories are shown up to the current frame by the colored curves. The different trajectories in the same spatial locations of the frame demonstrate the ability of our approach to capture the temporal motion variations of the crowd. For example, the green target in row 1 is moving in a completely different direction than the red and pink targets, although they share the spatial location where their trajectories intersect. Similarly, the pink, blue, red, and green targets in row 2 all move in different directions in the center part of the frame, yet our method is able to track each of these individuals. Such dynamic variations that we model using an HMM cannot be captured by a single motion model such as a "floor fields" [2]. Spatial variations are also handled by our approach, as illustrated by the targets concurrently moving in completely different directions in rows 5 and 6. In addition, our method is robust to partial occlusions as illustrated by the pink target in row 1, and the red targets in rows 3, 5, and 6.

Fig. 4.6 shows a failure case due to a severe occlusion. In these instances our method begins tracking the individual that caused the occlusion. This behavior, though not desired, shows the ability of our model to capture multiple motion patterns since the occluding individual is moving in a different direction. Other tracking failures occur due to poor texture. In the sidewalk scene, for example, the occasional viewable background and lack of texture on the pedestrians cause poorly-predicted local motion patterns. On such occasions, a local motion pattern that describes a relatively static structure, such as black clothing or the street, is predicted for a moving target. This produces non-smooth trajectories, such as the pink and red targets in row 5, or the red target in row 6 of Fig. 4.5.

Occasionally, an individual may move in a direction not captured by the training data. For instance, the pedestrian shown on the left of Fig. 4.7 is moving from left to right, a motion not

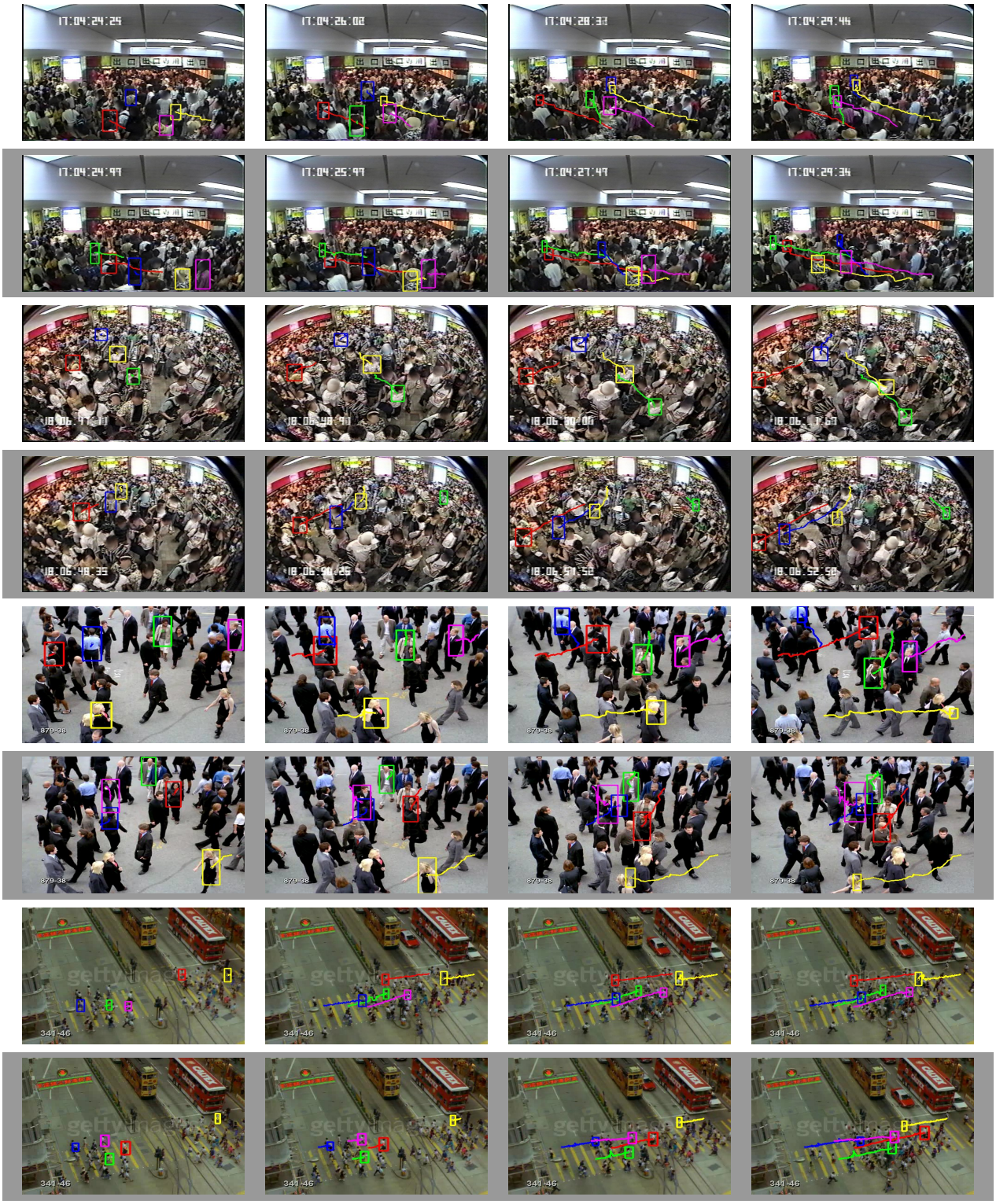


Figure 4.5: Frames showing our method of tracking pedestrians in videos crowded scenes.

Figure 4.6: Example of a tracking failure due to occlusion.

Figure 4.7: Tracking of pedestrians moving against the crowd.

present in the training data. Such cases are difficult to track since the space-time model can not predict the pedestrian's motion. On such occasions, the posteriors (given in Eq. 4.3) are near identical (since the emission probabilities are all close to 0), and thus the predicted optical flow is unreliable. This does not mean the targets can not be tracked, as shown by the correct trajectories in Fig. 4.7, but the tracking depends entirely on the appearance model. We address this shortcoming in Chap. 5 by measuring how much pedestrians deviate from the crowd.

We hand-labeled ground truth tracking results for 40 targets, 10 from each scene, to quantitatively evaluate our approach. Each target is tracked for at least 120 frames. The ground truth includes the target's position and the width and height of a bounding box. The concourse and ticket gate scenes contain many pedestrians whose lower bodies are not visible at all over the duration of the video. On such occasions, the ground truth boxes are set around the visible torso and head of the pedestrian. Given the ground truth state vector $\mathbf{k}_t$, we measure the error of the tracking result $\hat{\mathbf{k}}_t$ as $||\mathbf{k}_t - \hat{\mathbf{k}}_t||_2$.

Fig. 4.8 shows the error of our method for each labeled target, averaged over all of the frames in the video, compared to a particle filter using a color-histogram likelihood and second-degree

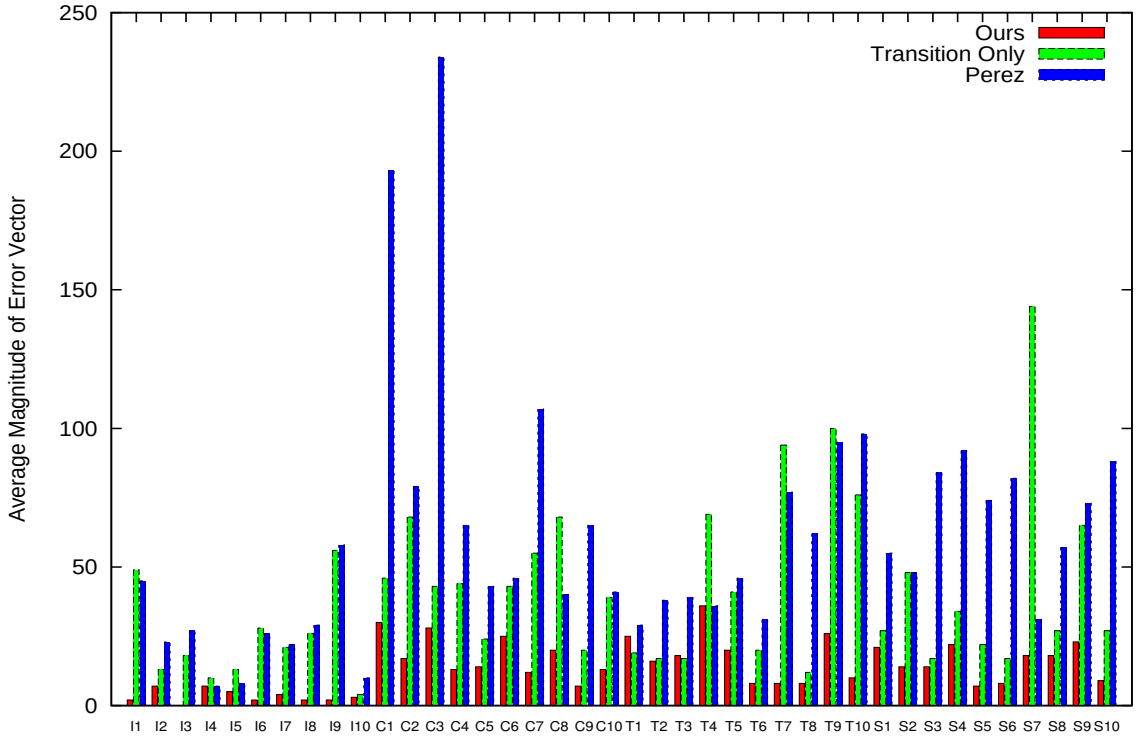


Figure 4.8: Tracking error using our approach compared with a second-degree auto-regressive model and using only our transition distribution with a color-based likelihood.

auto-regressive model [59] (labeled as Perez). In addition, we show the results using our predicted state-transition distribution with a color-histogram likelihood (labeled as Transition Only). On many of the targets our state transition distribution is superior to the second-degree autoregressive model, though 9 targets have a higher error. Our full approach improves the tracking results dramatically and consistently achieves a lower error than that of Pérez et al. [59].

Fig. 4.9 compares our approach with the "floor fields" method by Ali and Shah [2] and the topical model from Rodriguez et al. [62]. Since the other methods do not change the target's template size, we only measured the error in the x, y location of the target. Our approach more accurately tracks the pedestrian's locations in all but a few of the targets. The single motion model by Ali and Shah completely loses many targets that move in other directions. The method of Rodriguez et al. [62] models multiple possible movements, and thus achieves better results than the method of Ali and

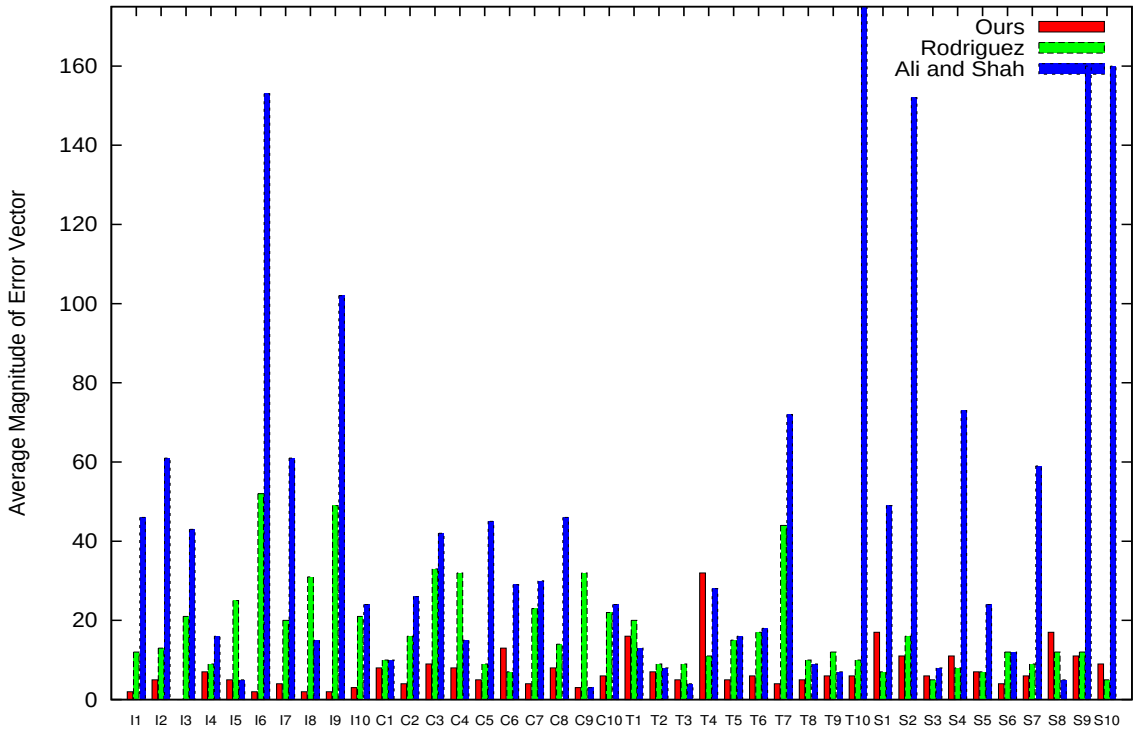


Figure 4.9: Tracking error of our approach compared with that of Ali and Shah [2], and Rodriguez et al. [62].

Shah, but is still limited since it does not include temporal information. Our temporally varying model allows us to track pedestrians in scenes that exhibit dramatic variations in the crowd motion.

Fig. 4.10 shows the tracking error over time, averaged over all of the targets, using our approach, that of Ali and Shah [2], and that of Rodriguez et al. [62]. The consistently lower error achieved by our approach indicates that we may track subjects more reliably over a larger number of frames. Our temporally varying model accounts for a larger amount of directional variation exhibited by the targets and enables accurate tracking over a longer period of time.

4.6.1 Using Gaussian Local Motion Patterns

Using the Gaussian representation of local motion patterns retains some appearance information, but reduces the accuracy of the predicted optical flow. The advantage of using the Gaussian repre-

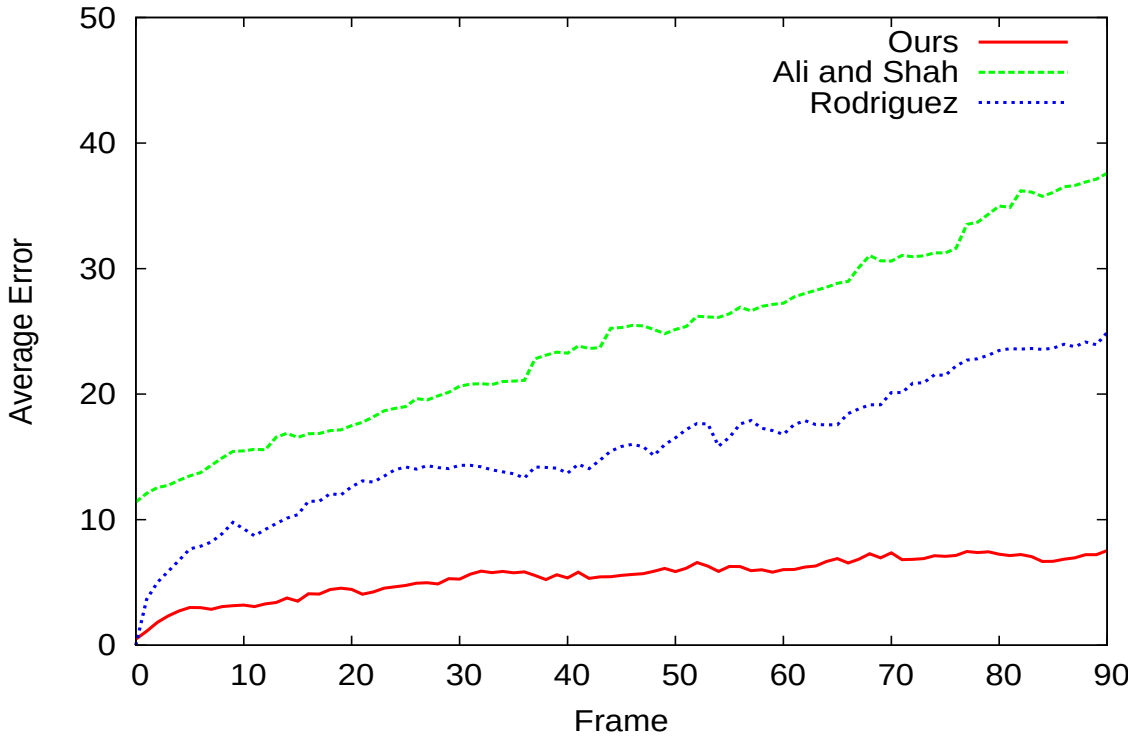


Figure 4.10: The average error of targets over time using our approach, that of Ali and Shah [2], and Rodriguez et al. [62].

sentation, however, is that we may dynamically adjust the likelihood distribution. Fig. 4.11 shows frames of a target during tracking and the variance of the likelihood distribution that we hypothesize from the predicted local motion patterns. The frames shown in the top row are indicated on the plot by blue boxes. When the target's appearance changes, for instance due to the partial occlusions on frames 56, 112, and 201, the variance is larger. The target's appearance has less variation in frames 29, 71, 154, and 171, and the hypothesized variance is also lower, allowing for more accurate tracking due to a narrower likelihood distribution.

4.7 Summary

In this chapter, we demonstrated how a model of the crowd motion may be leveraged as a scene-specific constraint on the motion of individuals. Specifically, we used the crowd motion to predict a

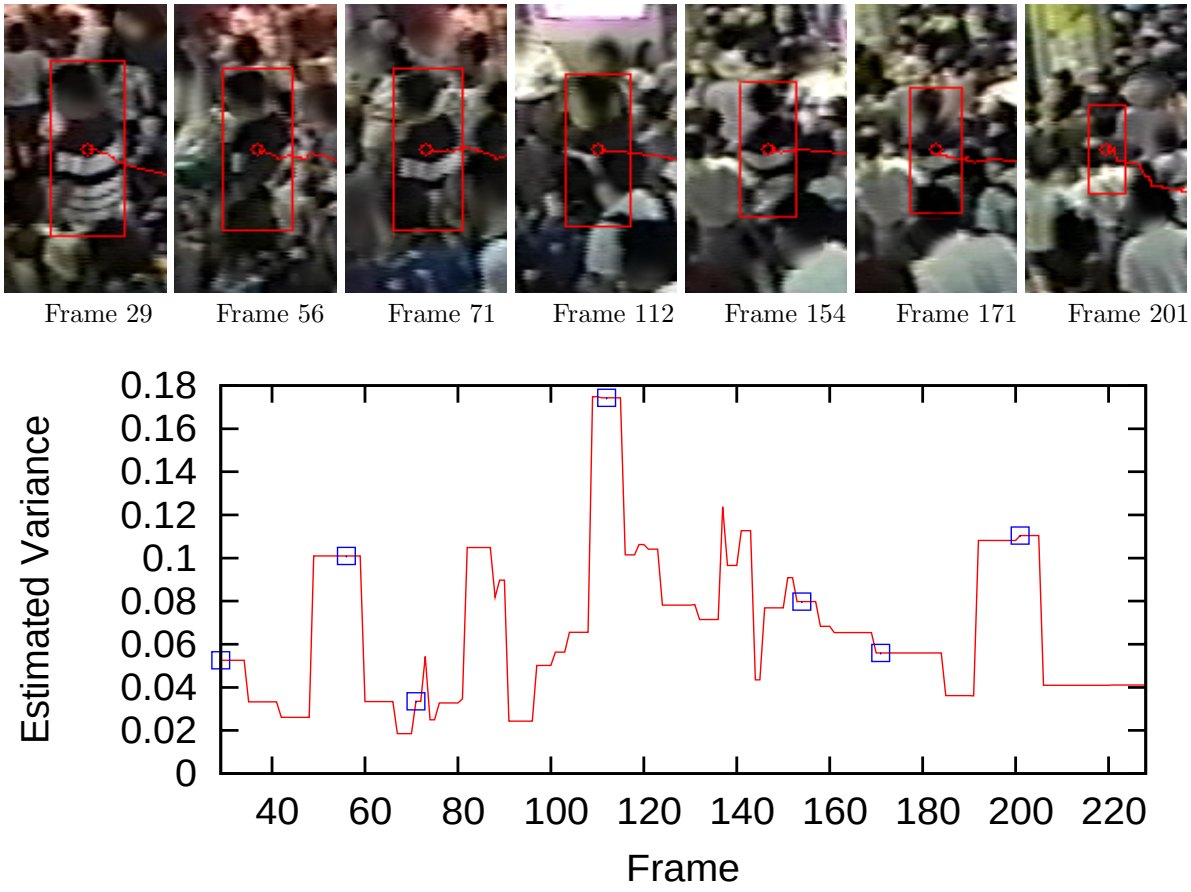


Figure 4.11: An illustration of our dynamically varying likelihood model.

local motion pattern at each space-time location of the video, which we used as a prior on a particle filter to track individuals. Since our model varies in the space-time, the prior on tracking changes depending on the video, vastly improving state-of-the-art tracking in crowded scenes.

Chapter 5: Deviations from the Crowd

5.1 Individuality

Previously we demonstrated how the scene-wide crowd motion can be leveraged for analysis of specific locations (Chap. 3) and individuals (Chap 4). In this chapter, we explore the relationship between the crowd and the individual pedestrians.

Up to this point, we have assumed that each the crowd motion is a smooth pattern formed by regularly occurring motions within the scene. Many crowd methods make simplistic assumptions about the individuals' motions, for instance, that they can be analytically expressed with hydrodynamics models [57]. Pedestrians, however, are individuals and do not necessarily follow such analytical rules. Individuals constantly anticipate and react to others surrounding them, causing pauses or changes in direction to avoid collisions. These interactions can be deviations from the crowd model and cause the performance of crowd-based methods to degrade.

Often pedestrians that deviate from the flow of the crowd are reacting to an interruption (e.g., someone cutting them off) or congestion. In such cases, the pedestrian can not move in the direction that then *intend* to move. Though knowing an individual's intention is impossible, the crowd motion encodes clues to where pedestrians wish to move. Still [68] notes that emergent behaviors (e.g., lanes or clusters) form in crowds because it is easier "to follow immediately behind someone who is already moving in your direction." In other words, the motion of the surrounding crowd often suggests *where* individuals intend to move.

In this chapter, we present a novel method for estimating how much a pedestrian conforms to the crowd. We the crowd motion to estimate the direction pedestrians will move in the future. We compare this "intended" motion to the instantaneous optical flow to measure *conformance*, i.e., how much the pedestrian at that space-time location is moving with the flow of the crowd. Space-time areas with low conformance contain pedestrians that deviate from the crowd, and cause crowd-based methods to fail. We demonstrate how conformance may be used to dynamically adjust our tracking

prior, enabling tracking of pedestrians who deviate from the crowd. In addition, we show how conformance is an indicator of unusual events in local areas and over the entire crowd.

5.2 Comparison with Related Work

Most crowd-based vision methods model the dynamics of the crowd as a collection of individuals obeying a set of analytical rules. Shah et al. [57], for example, treats each pedestrian as a particle in a fluid. As noted by Still [68], however, many emergent behaviors that exist in crowds do not occur in fluids. In contrast, we seek to measure how much each individual deviates from the crowd.

Mehran et al. [56] estimate the "social force," i.e., the influence of the crowd, on each individual. They use the instantaneous optical flow as intended velocity, and the average optical flow as the pedestrian's actual velocity. This assumption is not always valid: pedestrians tend to sway when their motion is restricted [6, 29], suggesting the instantaneous optical flow does not indicate their intended motion. As we show in Sec. 5.6, our method more accurately estimates the intended motion, and therefore better represents the deviations a pedestrian has from the crowd.

Bruan et al. [13] explore individuality in a crowd simulation method. Each individual agent is modeled using a measure of their altruism and their dependence on a group. Such characteristics, however, are studied only in the context of emergency events and crowd panic. They do not examine the impact of individuality on the motion of pedestrians in typical crowd scenarios, or how individuality plays a role in the formation of crowd dynamics.

It is worth noting the relationship of our measure to the "efficiency" presented by Helbing et al. [26]. They measure efficiency by comparing the intended motion of each pedestrian to their observed direction and velocity. Their measure, however, assumes that a pedestrian will not move faster than their desired speed (e.g., in a panic situation) and thus is not well bounded. In addition, they measure efficiency of each individual, rather than the image space measure that we propose. Finally, and perhaps most significantly, is that their measure is strictly theoretical and assumes the intention of each individual can be known. In contrast, our estimate of intention is based on the crowd motion, and thus is poor if a pedestrian is moving against the flow of the crowd. This is advantageous for our purpose: when pedestrians are moving against the crowd, their conformance

is low. In Helbing's model, such pedestrians may be moving efficiently but not conforming to the crowd flow.

5.3 Conformance

We start by defining how to measure conformance, assuming that the intentions of pedestrians are known. Sociologists Helbing and Vicsek [27] define the influence of surrounding pedestrians on an individual as the *interaction rate*, which is inversely related to conformance. Rather than computing how much the crowd influences each pedestrian, we estimate conformance at each space-time pixel location in the video. By doing so, we may analyze the scene without having to identify or track each pedestrian.

Let t be a time and $\mathbf{p} = [x, y]^T$ a 2D pixel location in the video. Let the intended motion of the pedestrian pedestrian (i.e., how they will move with the crowd) occupying pixel $\mathbf{p}$ at time t be defined by the 3D optical flow vector

$$\mathbf{u}_t(\mathbf{p}) = [\Delta x, \Delta y, \Delta t]^T \ , \tag{5.1}$$

where Δx, Δy, and Δt is the change in the horizontal, vertical, and temporal dimensions, respectively, and $|\mathbf{u}_t(\mathbf{p})|=1$. As in Sec. 2.3, diving by Δt yields the 2D optical flow

$$\tilde{\mathbf{u}}_t(\mathbf{p}) = [\Delta x/\Delta t, \Delta y/\Delta t] \, . \tag{5.2}$$

Similarly, let $\mathbf{v}_t(\mathbf{p})$ be the 3D instantaneous optical flow observed in the video.

We measure the deviation from the crowd flow by comparing the intended motion $\mathbf{u}_t(\mathbf{p})$ with the instantaneous flow $\mathbf{v}_t(\mathbf{p})$. As illustrated in Fig. 5.1(b), we use the great-circle distance to compute the deviation

$$d_t(\mathbf{p}) = \frac{\arccos\left(\mathbf{u}_t(\mathbf{p})^T \mathbf{v}_t(\mathbf{p})\right)}{\pi} \ . \tag{5.3}$$

Note that $d_t(\mathbf{p})$ is bounded by $[1, 0]$. We consider deviation inversely proportional to the conformance

(a) (b)

Figure 5.1: An example of measuring deviations from the crowd.

$$c_t(\mathbf{p}) = 1 - d_t(\mathbf{p}) \ . \tag{5.4}$$

Since we represent motion using 3D optical flow vectors, these equations capture both differences in direction (longitudinal variations across the unit sphere) and speed (latitudinal variations). To compute this, however, we need the intended motion $\mathbf{u}_t(\mathbf{p})$. Next, we describe how to use the crowd motion to estimate $\mathbf{u}_t(\mathbf{p})$.

5.4 Estimating Intended Motion

We use the trained HMMs to estimate the intended motion at each space-time location of a different video from the same scene. While it is impossible to know each individual's intention, pedestrians tend to *follow* others moving in the same intended direction [68]. As such, a pedestrian's future location within the scene is a reasonable indicator of where they intend to move, and how they will move with the crowd. We estimate the future location of pedestrians at each spatial point $\mathbf{p}$ from each time t in the video.

We use the HMMs to anticipate future flow fields, and approximate future locations by numeric integration. To anticipate future flow fields, we seek the posterior

$$\mathbf{y}_k(j) = \mathrm{p}(s_{t+k} = j | O_1, \ldots, O_t) \tag{5.5}$$

that represents the likelihood of being in state j and time $t+k$. In Sec. 2.4.1 we showed that given the observed local motions patterns $\{O_1, \ldots, O_t\}$ up to time t, the forwards step of the Forwards-Backwards algorithm can be used to compute the posterior

$$\hat{\boldsymbol{\alpha}}_t(j) = \mathrm{p}(s_t = j | O_1, \ldots, O_t) \ . \tag{5.6}$$

Using the transition matrix $\mathbf{A}$ of the HMM,

$$\mathbf{y}_k = \left[\hat{\boldsymbol{\alpha}}_t^T \mathbf{A}^k \right]^T \ . \tag{5.7}$$

Note that when $k = 1$, Eq. 5.7 is equivalent to the predictive posterior in Eq. 4.3. We compute K future densities $\{\mathbf{y}_k | k = 1 \ldots K\}$. As $k \to \infty$, Eq. 5.7 approaches the stationary distribution of the Markov process (if it exists). We select K large enough to approach the stationary distribution (typically $10 \leq K \leq 20$).

Let $\mathbf{f}_k$ be a 3D optical flow vector representing the motion of the crowd at time $k + t$. We anticipate the flow

$$\mathbf{f}_k = \mathbb{E}\left[\mathrm{p}(\boldsymbol{\mu}_{t+k} | O_1, \ldots, O_t) \right] \ , \tag{5.8}$$

where $\boldsymbol{\mu}_{t+k}$ is the mean direction of the emission densities. Via marginalization

$$\mathbf{f}_k = \sum_{j=1}^{J} \mathbb{E}\left[\mathrm{p}(\boldsymbol{\mu}_{t+k} | s_{t+k} = j) \right] \mathrm{p}(s_{t+k} = j | O_1, \ldots, O_t) \ . \tag{5.9}$$

Note that Eq. 5.9 is similar to the predicted local motion pattern used for tracking (Eq. 4.2), except that we are only considering the flow (i.e., we have dropped the concentration parameter) and have progressed k time instances instead of 1. The expected value $\mathbb{E}\left[\mathrm{p}(\boldsymbol{\mu}_{t+k} | s_{t+k} = j) \right]$ is simply the mean

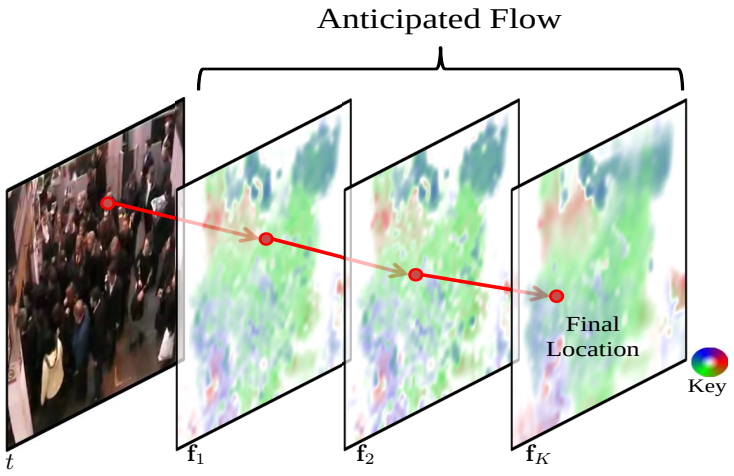


Figure 5.2: We estimate future location of each pixel by advancing it through a 3D flow field (color indicates speed and direction) anticipated by the crowd motion.

direction $\boldsymbol{\mu}_0^j$ of the hidden state j. Thus

$$\mathbf{f}_k = \frac{\sum_{j=1}^{J} \boldsymbol{\mu}_0^j \mathbf{y}_k(j)}{\left| \sum_{j=1}^{J} \boldsymbol{\mu}_0^j \mathbf{y}_k(j) \right|} , \tag{5.10}$$

where $\mathbf{y}_k$ is the future density from Eq. 5.7.

Let $\mathbf{f}_k(\mathbf{p})$ be the future flow vector computed at spatial location $\mathbf{p}$. We compute K future flow vectors for each spatial location $\mathbf{p} \in \mathcal{P}$ in the video volume. The resulting flow field $\{\mathbf{f}_k(\mathbf{p}) | \mathbf{p} \in \mathcal{P}, 1 \leq k \leq K\}$ represents the anticipated flow of the crowd.

As shown in Fig. 5.2, we estimate the future location of each point $\mathbf{p}$ starting at time t by advancing it through the anticipated flow field. Let $\tilde{\mathbf{f}}_k(\mathbf{p})$ be the 2D optical flow computed from $\mathbf{f}_k(\mathbf{p})$ (using the normalization in Eq. 5.2), and $\hat{\mathbf{p}}_k$ the location of $\mathbf{p}$ at time $k+t$. At $k = 0$, $\hat{\mathbf{p}}_0 = \mathbf{p}$. We advance the previous point to compute the next location

$$\hat{\mathbf{p}}_{k+1} = \hat{\mathbf{p}}_k + \tilde{\mathbf{f}}_k(\hat{\mathbf{p}}_k) . \tag{5.11}$$

The final location $\hat{\mathbf{p}}_K$ indicates, according the crowd motion, the future location of the pedestrian

occupying $\mathbf{p}$. Finally, we compute the intended direction

$$\overline{\mathbf{u}}_t(\mathbf{p}) = \hat{\mathbf{p}}_K - \mathbf{p} \tag{5.12}$$

and the intended motion

$$\mathbf{u}_t(\mathbf{p}) = \frac{1}{Z}\left[\overline{\mathbf{u}}_t(\mathbf{p})^T, K\right]^T \tag{5.13}$$

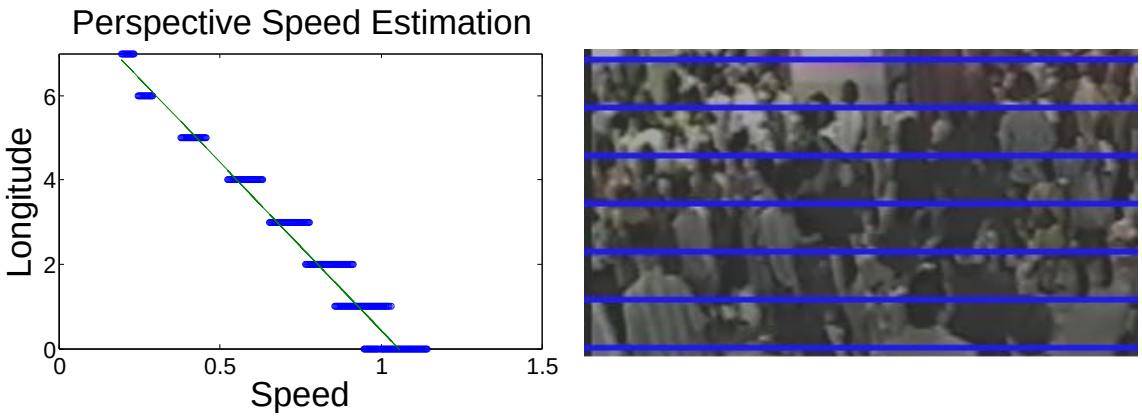
where Z is a normalization constant such that $|\mathbf{u}_t(\mathbf{p})| = 1$.

5.4.1 Intended Speed Without Congestion

Our estimate of the intended motion indicates where a pedestrian will to move according to our crowd model. Often, the crowd naturally moves slowly and thus the intended speed of the pedestrian is small. In the absence of a crowd, however, the natural walking speed of pedestrians is near constant. In some applications, knowing when pedestrians are *not* moving at their natural walking speed is desirable. We provide an estimate of pedestrian's natural walking speed, but note that in most applications Eq. 5.13 is correct. We demonstrate some cases of identifying pedestrians who are not moving in Sec. 5.6.

The walking speed of pedestrians has been well studied and is near constant if there is no congestion. Zip's [83] least-effort principle implies that pedestrians minimize metabolic energy when walking at roughly 1.33 meters per second [25], which has been verified in observational studies [71, 28]. For scenes recorded at a distance, we may assume orthographic camera projection and thus a constant intended speed can be estimated for all pedestrians. We approximate the intended speed as the maximum observed speed in the training video. Intuitively, we are identifying the few instances where pedestrians can move freely due to lulls in traffic or less-crowded areas. To address un-reliable or erroneous flow estimates, we use Chauvenet's criterion [18] to remove outliers.

As shown in Fig. 5.3, we estimate the intended speed in scenes with perspective projections by observing the relationship between each longitudinal frame location. First, we identify the fastest 5% of speeds measurements from each longitudinal frame location. Due to the perspective projection, the speeds across the frame have near-linear relationship. We use least-squares to fit a line to the



(a) (b)

Figure 5.3: Estimation of intended speed in perspective scenes.

speed measurements to estimate the desired speed over the entire image. Outliers are also removed using Chauvenet's criterion. Finally, given intended speed $s(\mathbf{p})$ and direction $\overline{\mathbf{u}}_t(\mathbf{p})$, we compute

$$\mathbf{u}_t(\mathbf{p}) = \frac{1}{Z}\left[\frac{\overline{\mathbf{u}}_t(\mathbf{p})^T}{|\overline{\mathbf{u}}_t(\mathbf{p})|}, s(\mathbf{p})\right]^T \tag{5.14}$$

where Z is a normalization constant such that $|\mathbf{u}_t(\mathbf{p})| = 1$.

5.5 Applications

5.5.1 Unusual Event Detection

The amount of pedestrians conforming to the crowd flow is a natural indicator of unusual activities. Atypical motions (i.e., pedestrians moving against the crowd) appear as deviations in local areas, and crowd disasters contain people moving irrationally. We identify *global* anomalies, i.e., affecting a large portion if not the entire crowd, by low values of the average conformance

$$\bar{c}_t = \frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}} c_t(\mathbf{p}) \tag{5.15}$$

where $\mathcal{P}$ is the set of 2D pixel locations.

Next, we consider the problem of detecting *local* anomalies, such as individuals moving against the

flow of the crowd. The conformance measure of such offenders will be low, and also cause deviations from the flow in their immediate vicinity (as surrounding pedestrians must avoid them). Since the steady-state of the crowd may differ between scenes, we compute the normalized conformance

$$\tilde{c}_t(\mathbf{p}) = \frac{c_t(\mathbf{p})}{Z(\mathbf{p})} \ , \tag{5.16}$$

where $Z(\mathbf{p})$ is the average conformance at spatial location $\mathbf{p}$ of the training data.

Let $\mathbf{p}_t = [\mathbf{p}, t]$ be the space-time point at pixel location $\mathbf{p}$ and time t. We identify the space-time locations with low values of $\tilde{c}_t(\mathbf{p})$ using a space-time Markov random field (STMRF). Specifically, we minimize the energy

$$\mathcal{E} = \sum_{t=1}^{T} \sum_{\mathbf{p}_t \in \mathcal{P}} \mathcal{V}(\mathbf{p}_t) + \sum_{\mathbf{y} \in \mathcal{N}(\mathbf{p}_t)} \mathcal{Q}(\mathbf{y}, \mathbf{p}_t) \ , \tag{5.17}$$

where $\mathcal{V}(\mathbf{p}_t)$ is the observation likelihood, $\mathcal{Q}(\mathbf{y}, \mathbf{p}_t)$ is a smoothing term, and $\mathcal{N}(\mathbf{p}_t)$ is the six space-time points in the neighborhood of the point $\mathbf{p}_t$. In our implementation, we use the library by Szeliski et al. [70] that implements graph-cuts [11, 12].

Let $g(\mathbf{p}_t)$ be the latent variable of the STMRF corresponding to the observation at point $\mathbf{p}_t$. Each latent variable has two possible values

$$g(\mathbf{p}_t) = \begin{cases} 0 & \text{if } \mathbf{p}_t \text{ is usual} \\ 1 & \text{otherwise.} \end{cases} \tag{5.18}$$

The observation likelihood is

$$\mathcal{V}(\mathbf{p}_t) = \begin{cases} \tilde{c}_t(\mathbf{p}) & \text{if } g(\mathbf{p}_t) = 1 \\ 1 - \tilde{c}_t(\mathbf{p}) & \text{otherwise.} \end{cases} \tag{5.19}$$

We use a scaled delta function for the smoothing term

$$\mathcal{Q}(\mathbf{y}, \mathbf{p}_t) = \begin{cases} 0 & \text{if } g(\mathbf{p}_t) = g(\mathbf{y}) \\ \sigma & \text{otherwise} \end{cases} \tag{5.20}$$

where σ is the smoothing variance selected empirically.

Minimizing the energy of the STMRF corresponds to identifying space-time regions with unusually high deviations. Such deviations occur when the motion of a pedestrian is not in the training data, i.e., an unusual event. Using our measure of conformance reduces many of the false positives compared with our method in Chap. 3, but does not operate online.

5.5.2 Tracking

We have estimated how much an individual is conforming to the flow of the crowd. As such, we may use it as a dynamic prior on the motion model in a particle filter for tracking pedestrians in crowded scenes.

Object-centric methods [59, 36] assume pedestrians exhibit smooth motion and impose (often first order) stochastic dynamics for the transition distribution $\mathrm{p}(\mathbf{k}_f|\mathbf{k}_{f-1})$ of the particle filter. In this spirit, we consider an *individual's* local motion pattern $\{\hat{\boldsymbol{\mu}}_f, \hat{\kappa}_f\}$ representing the motion from the actual tracked target up to time f. We learn this distribution online, i.e., it is updated after each tracking step. Specifically, at time f the previous locations of the target $\{\mathbf{k}_{f-1}, \ldots, \mathbf{k}_0\}$ are known. We use them to compute the observed motion

$$\mathbf{b}_{f-1} = \frac{1}{Z}\left[\left[\mathbf{k}_{f-1} - \mathbf{k}_{f-2}\right]^T, 1\right]^T , \tag{5.21}$$

where Z is a normalization constant such that $|\mathbf{b}_{f-1}| = 1$. We keep a running average of the observed motion

$$\overline{\mathbf{b}}_f = \alpha \mathbf{b}_{f-1} + (1-\alpha)\hat{\boldsymbol{\mu}}_{f-1} , \tag{5.22}$$

where α is a learning rate (we use 0.05). Next, we use the average to update the mean direction

$$\hat{\boldsymbol{\mu}}_f = \frac{1}{|\overline{\mathbf{b}}_f|}\overline{\mathbf{b}}_f \tag{5.23}$$

and the concentration parameter

$$\hat{\kappa}_f = \begin{cases} \dfrac{1}{(1 - |\overline{\mathbf{b}}_f|)} & \text{if } |\overline{\mathbf{b}}_f| \geq 0.9 \\ A_3^{-1}(|\overline{\mathbf{b}}_f|) & \text{otherwise,} \end{cases} \tag{5.24}$$

as in Sec. 2.3. Thus the von Mises-Fisher distribution defined by $\{\hat{\boldsymbol{\mu}}_f, \hat{\kappa}_f\}$ represents the motion of the individual being tracked.

Macroscopic crowd methods assume the crowd model yields an accurate prediction, and do not perform well when pedestrians deviate from the crowd (i.e., areas of low conformance). Object-centric methods struggle in areas without visible backgrounds (often high conformance). We use the conformance $c_f(\mathbf{p})$ as an indicator of how much to trust the crowd motion model. Recall from Sec. 4.4.1 that the transition distribution $\mathrm{p}(\mathbf{k}_f|\mathbf{k}_{f-1})$ of the particle filter is modeled using the predicted local motion pattern $\tilde{O}_t = \{\tilde{\boldsymbol{\mu}}_t, \tilde{\kappa}_t\}$. We combine this with the individual's local motion pattern $\{\hat{\boldsymbol{\mu}}_f, \hat{\kappa}_f\}$ using conformance as a weight to form a conformance-aware distribution $\{\boldsymbol{\mu}_f^d, \kappa_f^d\}$. Specifically

$$\kappa_f^d = (1 - c_f(\mathbf{k}_{f-1}))\, \hat{\kappa}_f + c_f(\mathbf{k}_{f-1})\tilde{\kappa}_t \tag{5.25}$$

and

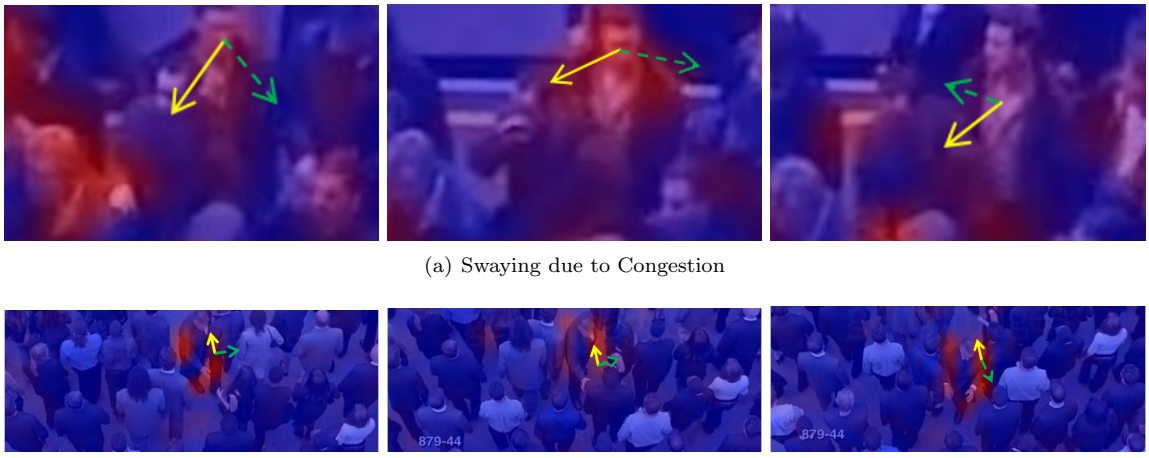
$$\boldsymbol{\mu}_f^d = \frac{1}{Z}\left[(1 - c_f(\mathbf{k}_{f-1}))\, \hat{\boldsymbol{\mu}}_f + c_f(\mathbf{k}_{f-1})\tilde{\boldsymbol{\mu}}_t\right] \tag{5.26}$$

where Z is a normalization constant such that $|\boldsymbol{\mu}_f^d| = 1$. The resulting distribution $\{\boldsymbol{\mu}_f^d, \kappa_f^d\}$ is sampled from in the transition step of the particle filter during tracking. To summarize, when $c_f(\mathbf{k}_{f-1})$ is high the crowd motion is used, and when $c_f(\mathbf{k}_{f-1})$ is low (indicating a deviation from the crowd) the individual's model is used.

5.6 Results

5.6.1 Examples of Deviations

Fig. 5.4 shows examples of pedestrians deviating from the crowd. Blue areas indicate high conformance, while red areas indicate deviations. The yellow solid arrow is intended motion, and the green



(a) Swaying due to Congestion

(b) Avoiding an Oncoming Individual

(c) Pedestrians Moving Slowly

Figure 5.4: Examples of pedestrians deviating from the crowd flow.

dashed arrow the optical flow. Fig. 5.4(a) shows an individual from the platform scene changing direction due to congestion. The individual's intended direction is to the left, and they deviate from the crowd when moving around other pedestrians. Fig. 5.4(b) shows pedestrians from the synthetic scene from [56] avoiding an oncoming individual. Their intended direction is vertical, and conformance decreases as they move to the side. In Fig. 5.4(c), from the concourse scene, we estimate the desired speed of pedestrians assuming the crowd is not present. As such, the pedestrians at the top of the frame are standing still and exhibit lower conformance than those moving in the lower left.

5.6.2 Accuracy

Since it is impossible to know a pedestrian's intentions, we cannot directly measure the accuracy of our estimated intended motion. We can, however, assume that pedestrians move in their intended direction over time. Let $\{\hat{\mathbf{k}}_t | t = 1 \ldots T\}$ be a sequence of ground-truth tracking locations for a specific

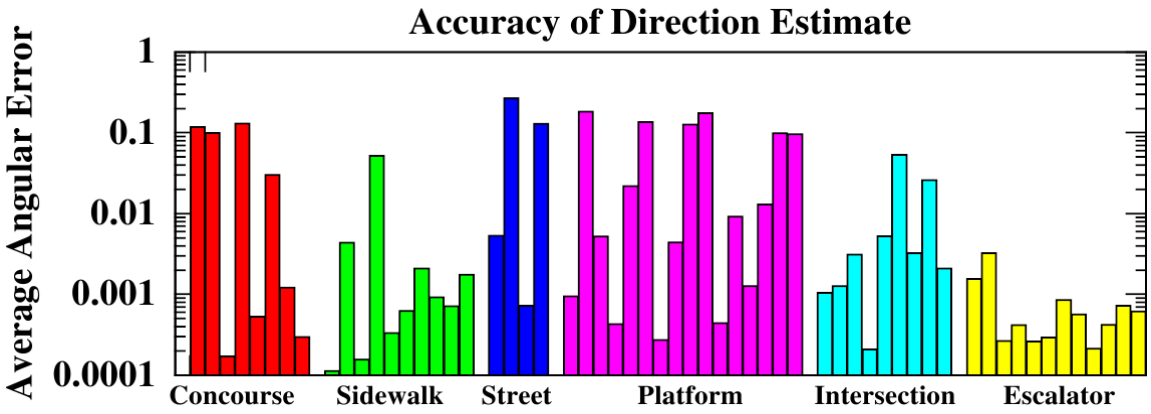


Figure 5.5: Angular error of our estimate of future direction for different pedestrians in six different scenes.

Table 5.1: The average accuracy of our estimate of future direction compared with Mehran et al. [56].

Scene	Concourse	Street	Sidewalk	Platform	Intersection	Escalator
Ours	0.042	0.086	0.006	0.048	0.010	0.001
Mehran et al. [56]	0.458	0.860	0.260	0.729	0.207	0.258

pedestrian. We measure the error

$$\frac{1}{T}\sum_{t=1}^{T}\arccos\left(\frac{\bar{\mathbf{u}}_t(\hat{\mathbf{k}}_t)^T\left[\hat{\mathbf{k}}_{t+w}-\hat{\mathbf{k}}_t\right]}{|\hat{\mathbf{k}}_{t+w}-\hat{\mathbf{k}}_t|}\right), \tag{5.27}$$

where $\bar{\mathbf{u}}_t$ is the estimated intended direction from Eq. 5.12, and w is a window size that depends on the subject (typically the duration the subject is in the scene).

Fig. 5.5 shows the estimation error for a number of subjects from six different scenes. Each bar indicates the error for a different individual. For almost all of the subjects the estimation error is below 0.1 (about 6°). None of the error rates exceed 0.2 which is small given the resolution of the video. The theoretical maximum error is π, and thus at most the error is $0.2/\pi \approx 6\%$.

Table 5.1 shows the average error for all scenes, and the error using the optical flow for the intended motion as suggested by Mehran et al. [56]. Scenes with less structure, such as the concourse or street, have higher errors due to the larger number of directions that pedestrians move. Compared with Mehran et al. [56], our method achieves consistently lower errors.

(a) (b) (c)

Figure 5.6: Visualization of conformance for the three scenes from the UMN dataset [73].

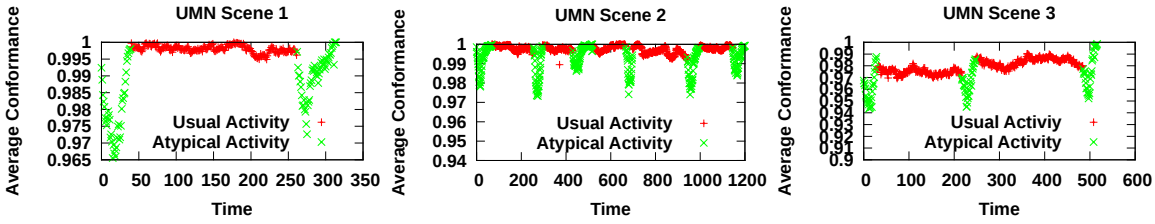


Figure 5.7: The average conformance plotted over time for the three scenes in the UMN dataset [73].

5.6.3 Impact on Unusual Event Detection

First, we detect global anomalies as frames with low average conformance on the University of Minnesota Crowd Dataset [73]. The dataset contains a number of usual and unusual video segments from 3 different scenes. For each scene, we train the HMMs on a usual sequence, and estimate conformance on the remaining sequences. A frame is considered unusual if its average conformance is below a specific threshold that is selected empirically. Fig. 5.6 shows visualizations of the conformance for usual (top) and unusual activity (bottom) for all three scenes. The pedestrians in the unusual frame (bottom) exhibit lower conformance than those in the usual frame (top).

Fig. 5.7 shows the average conformance plotted over time for each scene in the UMN data set.

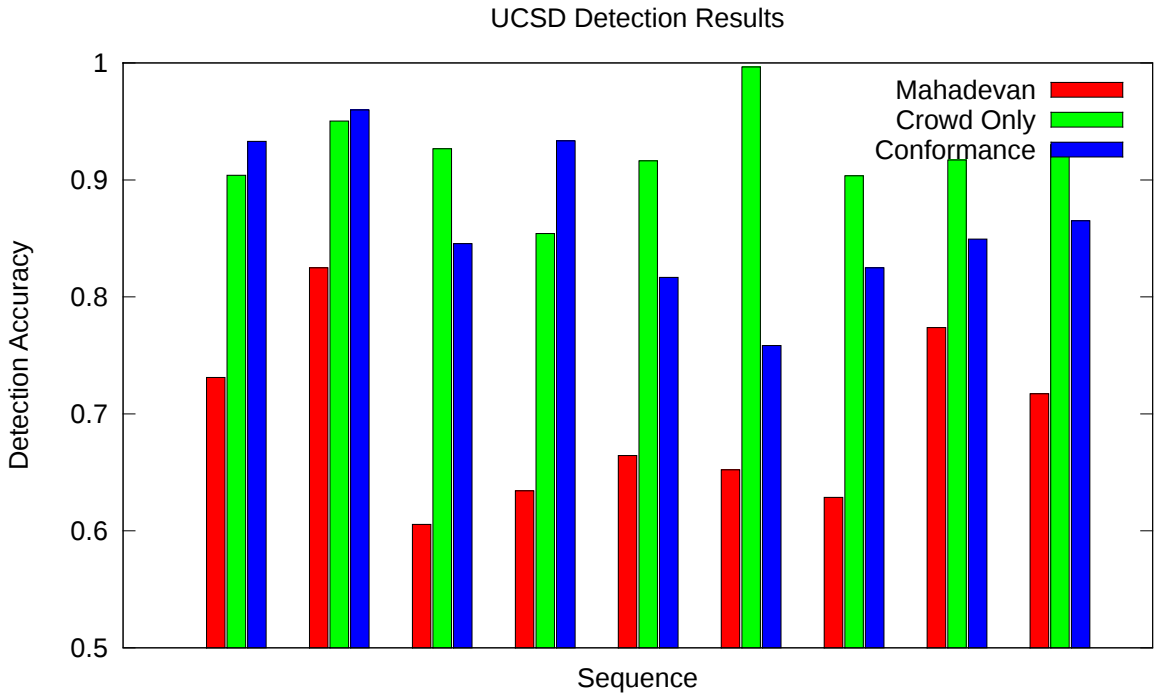


Figure 5.8: Detection accuracy on the UCSD dataset using conformance, our method from Chap. 3, and that of Mahadevan et al. [51].

The red points correspond to frames from usual activity, and the green points to unusual activity. The conformance clearly drops during clips of unusual activity. We vary the threshold to compute an ROC curve whose area is 0.92, compared with 0.96 in [56] and 0.99 in [61]. Our poorer performance is due to the higher conformance at the beginning and end of each unusual sequence (where pedestrians are moving normally) as visible in each graph in Fig. 5.7.

Fig. 5.8 shows the detection accuracy using conformance for 9 sequences compared with Mahadevan et al. [51] and our previous approach from Chap. 3 (Crowd Only). We use the results of Mahadevan et al. [51] posted on the web [52] for comparison. Using conformance to detect unusual events yields a higher accuracy rate than that of Mahadevan et al. [51], but only occasionally out-performs our previous method.

Though conformance does not out-perform our previous method, it may be used to achieve a more accurate localization of unusual events. Fig. 5.9 shows detection results on the concourse clip using conformance and our previous approach from Chap. 3. Here, we use a smaller cuboid size

(a) (b) (c)

Figure 5.9: Local unusual event detection (a) using conformance (b) compared with our previous approach (c).

$(20 \times 20 \times 5)$ to better localize the unusual event. As visible in Fig. 5.9(c), our previous approach has some errors. Fig. 5.9(b) shows the conformance, which is clearly lower in the general area of the offending pedestrians. As a result, the results of our conformance-based method (Fig. 5.9(c)) has fewer errors.

5.6.4 Impact on Tracking

We quantitatively evaluate our tracking method using hand-labeled ground truth of targets. Given a ground-truth location $\hat{\mathbf{k}}_t$ and tracking result $\mathbf{k}_t$, the tracking error is $|\hat{\mathbf{k}}_t - \mathbf{k}_t|$ averaged over all of the frames $\{t=1,\ldots,T\}$. Fig. 5.10 shows the tracking error for a number of subjects using conformance with our crowd-only method from Chap. 4 and that of Rodriguez et al. [62]. Table 5.2 shows the tracking error averaged over all pedestrians for each scene. With few exceptions, including

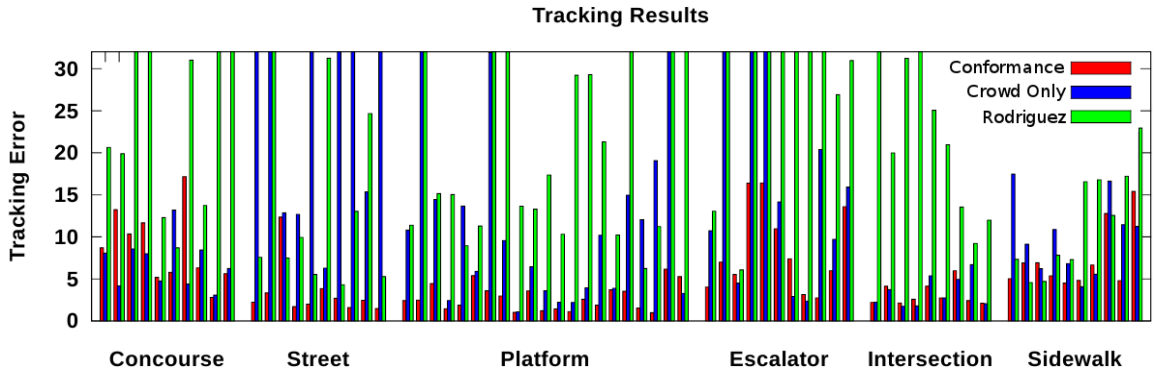


Figure 5.10: Tracking error for multiple subjects using divergence compared with our crowd-only model and that of Rodriguez et al. [62].

Table 5.2: The average tracking error of pedestrians in different scenes using divergence, our crowd-only model, and the method of Rodriguez et al. [62].

	Concourse	Street	Platform	Escalator	Intersection	Sidewalk
With Conformance	8.7	**3.3**	**2.8**	**8.5**	**3.1**	**7.4**
Crowd Only Model	**6.8**	47.6	17.3	24.8	3.56	9.9
Rodriguez et al. [62]	24.7	14.8	29.9	60.4	25.9	11.9

conformance in tracking yields comparable results on many scenes, and significantly lower error on the platform, escalator, and street scenes where pedestrians deviate from the structured environment.

Pedestrians that deviate from the flow of the crowd present challenges to tracking. Since such pedestrians naturally have low conformance, our method is able to track them. Fig. 5.11 shows two tracking results using our conformance-based tracking and our crowd-only model. In both cases, the pedestrian is moving against the crowd: the first is moving left to right, and the second is moving towards the bottom of the frame. As shown in green, the crowd model assumes pedestrians are moving with the crowd, drifts, and loses the target. The conformance-based method, shown in red, is able to compensate for the anomaly and accurately track the targets. Fig. 5.12 shows the tracking error for 16 anomalous targets using both methods. Using conformance achieves a consistently lower error.

Fig. 5.13 shows the ratio of the tracking error using conformance compared with the error using our method from Chap. 4 for different subjects. A high ratio (i.e., 1) indicates the errors are near identical and adding conformance yields no improvement, while a low ratio indicates a high improvement. The downward trend of the points is intuitive: using conformance to adjust the tracking prior

Figure 5.11: Tracking results of anomalous pedestrians from the concourse (top) and UCSD dataset (bottom) using conformance (red) and a crowd model (green).

vastly improves results when pedestrians deviate from the crowd, and performs similarly to crowd models when pedestrians are moving with the flow.

5.7 Summary

In this chapter, we presented an automatic method for estimating the amount pedestrians deviate from the learned crowd model. We demonstrated that the crowd motion can be leveraged to accurately estimate the direction that pedestrians will travel over time. Comparing this estimate to the observed optical flow yielded a measure of conformance to the crowd flow. By identifying low values of conformance, we detected unusual crowd behavior on the global and local scale. In addition, we demonstrated how conformance could be incorporated into our tracking method to handle cases of pedestrians deviating from the crowd.

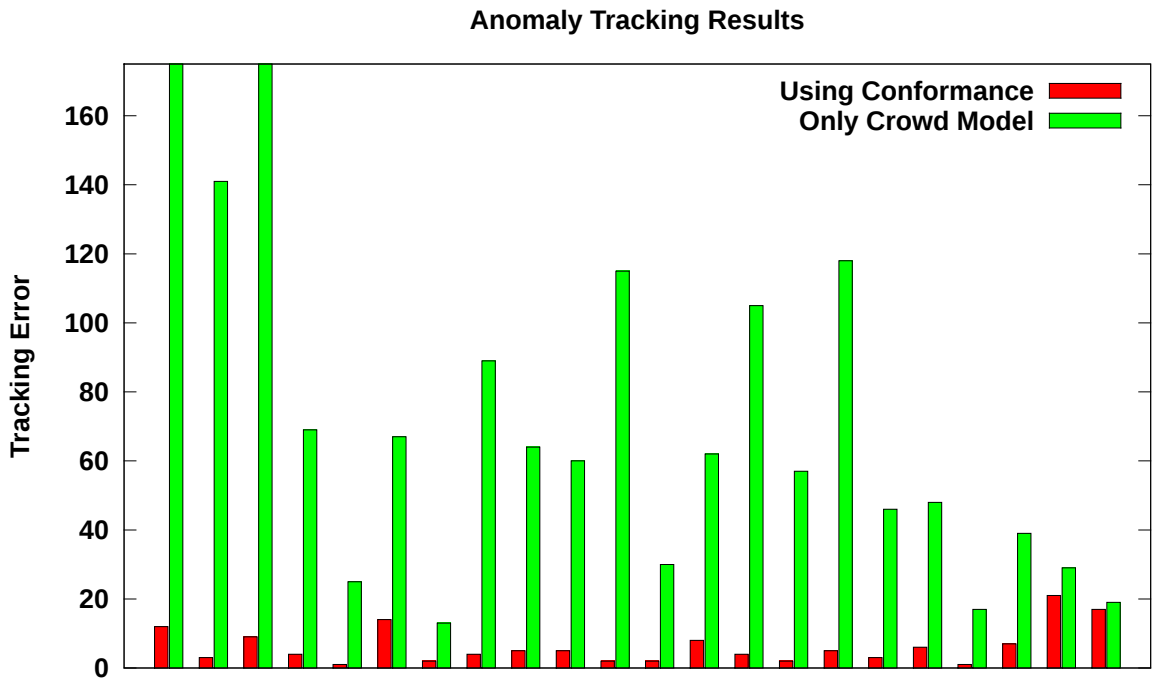


Figure 5.12: Error in tracking anomalous pedestrians using only the crowd model and our measure of conformance.

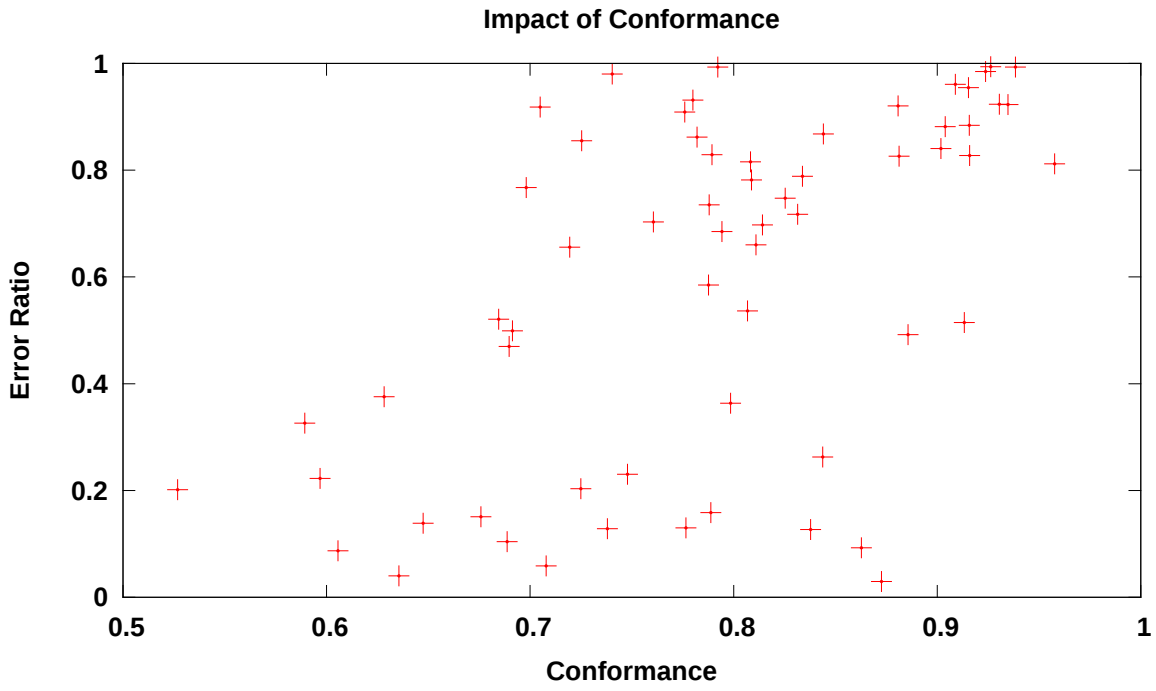


Figure 5.13: Improvement of tracking results for subjects that exhibit different levels of conformance.

Chapter 6: Conclusions

6.1 Summary

In this dissertation, we have explored how to learn the underlying structured pattern within videos of crowded scenes, and how it may be leveraged as a *scene-centric* constraint for automatic video analysis. To achieve this, we presented a novel statistical model based on local motion patterns that represents the spatially and temporally dynamic nature of the crowd motion. We then explored what information is encoded in our model while addressing three problems facing vision-based analysis of crowded scenes. In each application, our framework enables accurate and robust analysis of videos containing crowd activity too complex for conventional methods.

We first presented a method for learning the underlying structured pattern of the crowd. We captured the characteristic spatial and temporal variations of motion in the crowd by training a collection of hidden Markov models over local motion patterns. Each local motion pattern is a directional distribution of optical flow vectors, encoding the uncertainty in motion that occurs in real-world videos due to poor texture or complex motion. This holistic method captures the space-time nature of the crowd without without identifying, modeling, or counting the pedestrians within the scene.

When applied to unusual event detection, our model encodes the *steady-state*, i.e., typical behavior, of the crowd. Thus, we detected unusual events by identifying local motion patterns that statistically deviated from the learned model. This approach does not rely on knowing each possible activity ahead of time, only the spatio-temporal variations that comprise the "usual" crowd activity. The results show our method accurately detects local unusual events in videos of complex, crowded scenes that contain a significant number of pedestrians. A primary advantage of our method is the ability to detected subtle events occurring in specific locations within the scene. Such detection is of practical use in real-world surveillance applications that need fine-grained analysis beyond detecting crowd disasters. In addition, it can be tailored to a specific scene or activity that the surveillance

operator wishes to analyze.

Next, we demonstrated that the crowd motion could be leveraged to predict the movement of a single pedestrian, and we used this information as a prior for tracking individuals. Since our model encoded the temporal statistics between local motion patterns, we accurately predicted distributions of optical flow at each space-time point in the video. This predicted flow served as a prior on a particle filter that changed dynamically over the space-time video volume. Using a spatially and temporally changing prior enabled state of the art tracking results on scenes that cause other crowd-based methods to fail.

Finally, we leveraged our model to measure how much pedestrians conform to the flow of the crowd. We demonstrated that the crowd motion encodes clues to the future locations of pedestrians, which we used as an estimate of where they intend to move. By comparing this estimate to the instantaneous optical flow, we measured the conformance of pedestrians throughout the video. This measure represents the impact that individuality can have on the motion of the crowd. In this spirit, we demonstrated how to further improve tracking results for pedestrians that deviate from the crowd model. Experimental results demonstrate accurate and robust tracking may be achieved on complex scenes where crowd-only or object-only methods degrade. Additionally, we showed how deviations from the crowd correspond to both local and global unusual events. This work bridges the gap between crowd-centric and object-centric methods, and provides a means to study the relationship between individuals and the crowd that they comprise.

6.2 Future Work

We conclude with a brief discussion of future work.

Crowd disasters are a primary motivation for the study of crowd dynamics and crowd simulation. In this dissertation, we have advanced the field of automatic, vision-based crowd analysis. It remains to be seen, however, how this work can impact the inverse problem of crowd simulation especially with regards to evacuation dynamics. One immediate application of our crowd model could be the estimation of parameters for simulation techniques in an effort to make them more realistic. In addition, the cooperation between vision and simulation has the potential to impact real-world

scenarios. Crisis intervention, for example, may be viewed as a two step process: use vision to observe the current crowd dynamics and simulation to anticipate if a crisis is likely to occur. After anticipating crisis situations, disasters can be avoided by changing the environment or redirecting the crowd flow.

In this dissertation, we have explored crowded scenes of *pedestrians* which naturally exhibit an underlying latent structure. Many of our assumptions are based on the fact that we are analyzing humans. We believe that our basic approach (i.e., learning statistical models over local motion patterns) has applications in other scenes that contain large numbers of constituents and dynamically varying patterns. Many forms of wildlife, for example, flock together and form collective, global structures. While biologists have studied these behaviors, few if any have used vision techniques. By automatically analyzing the behavior of wildlife we may achieve a greater understanding of behavior, migration patterns, and social structures. We believe that our latent models, learned from real-world complex videos, can play a pivotal role in analyzing the complex behaviors of such animals.

Finally, we believe our crowd methods have the potential to impact autonomous navigation systems in public areas. Navigation planning through public areas has vast applications for assisting the vision and mobility impaired as well as advancing the current state of autonomous robots. Security cameras continuously capture the state of the environment and may be used to augment already existing navigation techniques. We believe our crowd analysis framework can play a key role in monitoring and anticipating the motion of pedestrians in crowded areas, enabling such systems to seamlessly and efficiently navigate through the crowd.

Bibliography

[1] S. Ali and M. Shah. A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[2] S. Ali and M. Shah. Floor Fields for Tracking in High Density Crowd Scenes. In *Proc. of European Conf on Computer Vision*, 2008.

[3] E. Andrade, S. Blunsden, and R. Fisher. Modelling Crowd Scenes for Event Detection. In *Proc. of International Conf on Pattern Recognition*, pages 175–178, 2006.

[4] E. L. Andrade, S. Blunsden, and R. B. Fisher. Hidden markov models for optical flow analysis in crowds. In *Proc. of International Conf on Pattern Recognition*, pages 460–463, 2006.

[5] S. M. Arulampalam, S. Maskell, and N. Gordon. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002.

[6] C. B. B. Krausz. Analyzing Pedestrian Behavior in Crowds for Automatic Detection of Congestions. In *ICCV-Workshop on Modeling, Simulation an Visual Analysis of Large Crowds*, 2011.

[7] M. Bangert, P. Hennig, and U. Oelfke. Using An Infinite Von Mises-Fisher Mixture Model to Cluster Treatment Beam Directions in External Radiation Therapy. In *International Conference on Machine Learning and Applications*, 2010.

[8] M. Betke, D. Hirsh, A. Bagchi, N. Hristov, N. Makris, and T. Kunz. Tracking Large Variable Numbers of Objects in Clutter. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, October 2007.

[10] M. J. Black and D. J. Fleet. Probabilistic Detection and Tracking of Motion Boundaries. *Int'l Journal on Computer Vision*, 38(3):231–245, July 2000.

[11] Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, Sep. 2004.

[12] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, Nov. 2001.

[13] A. Braun, S. R. Musse, L. P. L. De Oliveira, and B. E. J. Bodmann. *Modeling individual behaviors in crowd simulation*, pages 143–148. IEEE Comput. Soc, 2003.

[14] G. Brostow and R. Cipolla. Unsupervised Bayesian Detection of Independent Motion in Crowds. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 594–601, June 2006.

[15] A. B. Chan and N. Vasconcelos. Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(5):909–26, May 2008.

[16] M. T. Chan, A. Hoogs, J. Schmiederer, and M. Petersen. Detecting Rare Events in Video Using Semantic Primitives With HMM. In *Proc. of International Conf on Pattern Recognition*, pages 150–154, 2004.

[17] U. Chattaraj, A. Seyfried, and P. Chakroborty. Comparison of Pedestrian Fundamental Diagram Across Cultures. *Advances in Complex Systems*, 12(03):393–405, 2009.

[18] W. Chauvenet. *A Manual of Spherical and Practical Astronomy*, pages 474–566. Adamant Media Corporation, 5 edition, 1891.

[19] A. Cheriyadat and R. Radke. Detecting Dominant Motions in Dense Crowds. *Selected Topics in Signal Processing, IEEE Journal of*, 2(4):568–581, 2008.

[20] S. C. Choi and R. Wette. Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, 11(4):683–690, 1969.

[21] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, 2005.

[22] H. Dee and D. Hogg. Detecting Inexplicable Behaviour. In *Proc. of British Macine Vision Conf*, pages 477–486, 2004.

[23] A. Gilbert and R. Bowden. Multi Person Tracking Within Crowded Scenes. In *IEEE Workshop on Human Motion*, pages 166–179, 2007.

[24] J. Gryn, R. Wildes, and J. Tsotsos. Detecting Motion Patterns via Direction Maps With Application to Surveillance. In *IEEE Workshop on Motion and Video Computing*, pages 202–209, 2005.

[25] S. Guy, J. Chhugani, S. Curtis, P. Dubey, M. Lin, and D. Manocha. PLEdestrians : A Least-Effort Approach to Crowd Simulation. In *Proc. of Eurographics/ACM SIGGRAPH Symposium on Computer Animation*, pages 119–128, 2010.

[26] D. Helbing, Moln, I. J. Farkas, and K. Bolay. Self-Organizing Pedestrian Movement. *Environment and Planning B: Planning and Design*, 28(3):361–383, 2001.

[27] D. Helbing and T. Vicsek. Optimal Self-Organization. *New Journal of Physics*, 13, 1999.

[28] L. F. Henderson. The Statistics of Crowd Fluids. *Nature*, 229, 1971.

[29] S. P. Hoogendoorn and W. Daamen. Pedestrian Behavior at Bottlenecks. *Transportation Science*, 39, 2005.

[30] B. K. P. Horn and B. G. Schunck. Determining Optical Flow. Technical report, Cambridge, MA, USA, 1980.

[31] T. Hospedales, S. Gong, and T. Xiang. A Markov Clustering Topic Model for Mining Behaviour in Video. In *Proc. of IEEE Int'l Conf on Computer Vision*, 2009.

[32] M. Hu, S. Ali, and M. Shah. Detecting global motion patterns in complex videos. In *Proc. of International Conf on Pattern Recognition*, 2008.

[33] M. Hu, S. Ali, and M. Shah. Learning motion patterns in crowded scenes using motion flow field. In *Proc. of International Conf on Pattern Recognition*, pages 1–5, 2008.

[34] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A System for Learning Statistical Motion Patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(9):1450–1464, Sep. 2006.

[35] C. Hue, J.-P. Le Cadre, and P. Perez. Posterior Cramer-Rao Bounds for Multi-Target Tracking. *Aerospace and Electronic Systems, IEEE Transactions on*, 42(1):37 – 49, Jan. 2006.

[36] M. Isard and A. Blake. CONDENSATION-Conditional Density Propagation for Visual Tracking. *Int'l Journal on Computer Vision*, 29(1):5–28, August 1998.

[37] N. Johnson and D. Hogg. Learning the Distribution of Object Trajectories for Event Recognition. In *Proc. of British Macine Vision Conf*, pages 583–592, 1995.

[38] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Crowded Videos. In *Proc. of IEEE Int'l Conf on Computer Vision*, pages 1–8, 2007.

[39] Z. Khan, T. Balch, and F. Dellaert. MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(11):1805–1819, Nov. 2005.

[40] Z. Khan, T. Balch, and F. Dellaert. MCMC Data Association and Sparse Factorization Updating for Real Time Multitarget Tracking with Merged and Multiple Measurements. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):1960–1972, Oct. 2006.

[41] A. Kläser, M. Marszałek, and C. Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In *Proc. of British Macine Vision Conf*, pages 995–1004, 2008.

[42] L. Kratz and K. Nishino. Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 1446–1453, 2009.

[43] L. Kratz and K. Nishino. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:987–1002, 2012.

[44] T. Kretz, A. Grunebohm, M. Kaufman, F. Mazur, and M. Schreckenberg. Experimental Study of Pedestrian Counterflow in a Corridor. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(10):P10001, 2006.

[45] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[46] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, 2008.

[47] B. Leibe, E. Seemann, , and B. Schiele. Pedestrian Detection in Crowded Scenes. June 2005.

[48] J. Li, S. Gong, and T. Xiang. Global Behaviour Inference Using Probabilistic Latent Semantic Analysis. In *Proc. of British Macine Vision Conf*, 2008.

[49] J. Li, S. Gong, and T. Xiang. Scene Segmentation for Behaviour Correlation. In *Proc. of European Conf on Computer Vision*, pages 383–395, 2008.

[50] Y. Li, C. Huang, and R. Nevatia. Learning to Associate: HybridBoosted Multi-Target Tracker for Crowded Scene. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, 2009.

[51] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly Detection in Crowded Scenes. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010.

[52] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Results of MDT based Anomaly Detection and Localization in Crowded Scenes. http://www.svcl.ucsd.edu/projects/anomaly/results.html, 2010.

[53] A. Mardia and S. El-Atoum. Bayesian Inference for The Von Mises-Fisher Distribution Miscellanea. *Biometrika*, 63(1):203–206, 1976.

[54] K. V. Mardia and P. Jupp. *Directional Statistics*. John Wiley and Sons Ltd, 1999.

[55] R. Mehran, B. E. Moore, and M. Shah. A streakline representation of flow in crowded scenes. In *European Conference on Computer Vision(ECCV)*, 2010.

[56] R. Mehran, A. Oyama, and M. Shah. Abnormal Crowd Behavior Detection using Social Force Model. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, 2009.

[57] B. E. Moore, S. Ali, R. Mehran, and M. Shah. Visual Crowd Surveillance Through a Hydrodynamics Lens. *Commun. ACM*, 54:64–73, 2011.

[58] O. Nestares and D. J. Fleet. Probabilistic Tracking of Motion Boundaries with Spatiotemporal Predictions. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 358–365, 2001.

[59] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-Based Probabilistic Tracking. In *Proc. of European Conf on Computer Vision*, pages 661–675, 2002.

[60] L. Rabiner. A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77(2):257–286, Feb. 1989.

[61] R. Raghavendra, A. D. Bue, M. Cristani, and V. Murino. Optimizing Interaction Force for Global Anomaly Detection in Crowded Scenes. In *Proc. of IEEE Int'l Conf on Computer Vision*, pages 136–143, 2011.

[62] M. Rodriguez, S. Ali, and T. Kanade. Tracking in Unstructured Crowded Scenes. In *Proc. of IEEE Int'l Conf on Computer Vision*, 2009.

[63] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert. Density-aware person detection and tracking in crowds. In *Proc. of IEEE Int'l Conf on Computer Vision*, 2011.

[64] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[65] M. D. Rodriguez and M. Shah. Detecting and segmenting humans in crowded scenes. In *Proceedings of the 15th international conference on Multimedia*, pages 353–356, 2007.

[66] A. Schadschneider, W. Klingsch, H. Kluepfel, T. Kretz, C. Rogsch, and A. Seyfried. Evacuation Dynamics: Empirical Results, Modeling and Applications. *Encyclopedia of Complexity and Systems Science*, pages 3142–3176, 2009.

[67] E. Shechtman and M. Irani. Space-Time Behavior Based Correlation. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 405–412, 2005.

[68] K. Still. *Crowd Dynamics*. PhD thesis, University of Warwick, 2000.

[69] D. Sugimura, K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Using Individuality to Track Individuals: Clustering Individual Trajectories in Crowds Using Local Appearance and Frequency Trait. In *Proc. of IEEE Int'l Conf on Computer Vision*, 2009.

[70] R. Szeliski, R. Zabih, D. Scharstein, V. K. O. Veksler, A. Agarwala, M. Tappen, and C. Rother. A Comparative Study of Energy Minimization Methods for Markov Random Fields. In *Proc. of European Conf on Computer Vision*, 2006.

[71] K. Teknomo. *Microscopic Pedestrian Flow Characteristics: Development of an Image Processing Data Collection and Simulation Model*. PhD thesis, Tohoku University, 2002.

[72] University of California San Diego. Anomaly Detection Dataset. http://www.svcl.ucsd.edu/projects/anomaly/, 2010.

[73] University of Minnesota. Unusual Crowd Activity Dataset. http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi, 2006.

[74] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31:539–55, March 2009.

[75] X. Wang, K. Tieu, and E. Grimson. Learning Semantic Scene Models By Trajectory Analysis. In *Proc. of European Conf on Computer Vision*, pages 110–123, 2006.

[76] J. Wright and R. Pless. Analysis of Persistent Motion Patterns Using the 3D Structure Tensor. In *IEEE Workshop on Motion and Video Computing*, pages 14–19, 2005.

[77] B. Wu and R. Nevatia. Tracking of Multiple, Partially Occluded Humans Based On Static Body Part Detection. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 951–958, 2006.

[78] Y. Yang, J. Liu, and M. Shah. Video Scene Understanding Using Multi-scale Analysis. In *Proc. of IEEE Int'l Conf on Computer Vision*, 2009.

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

[80] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *Mach. Vision Appl.*, 19(5):345–357, 2008.

[81] T. Zhao, R. Nevatia, and B. Wu. Segmentation and Tracking of Multiple Humans in Crowded Environments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(7):1198–1212, 2008.

[82] S. Zhou, D. Chen, W. Cai, L. Luo, M. Y. H. Low, F. Tian, V. S.-H. Tay, D. W. S. Ong, and B. D. Hamilton. Crowd modeling and simulation technologies. *ACM Trans. Model. Comput. Simul.*, 20(4):1–35, 2010.

[83] G. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley Press, 1949.

Vita

Louis Aloysious Kratz III received his BS and MS degrees in Computer Science from Drexel University in 2006, graduating *Magna Cum Laude*.

Journal Papers

- *Tracking Pedestrians using Local Spatio-temporal Motion Patterns in Extremely Crowded Scenes* Louis Kratz and Ko Nishino. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34:987–1002, 2012.
- *Bayesian Defogging* Ko Nishino, Louis Kratz, and Stephen Lombardi. International Journal of Computer Vision, 98:263–278, 2012.

Book Chapters

- *Spatio-Temporal Motion Pattern Models of Extremely Crowded Scenes* Louis Kratz and Ko Nishino. L. Wang, G. Zhao, L. Cheng, and M. Pietikainen, editors. In Machine Learning for Vision-Based Motion Analysis. Springer, 2011.

Selected Conference Papers

- *Making Gestural Input from Arm-Worn Inertial Sensors More Practical* Louis Kratz, Scott Saponas, and Dan Morris. ACM SIGCHI Conference on Human Factors in Computing Systems, 2012.
- *Factorizing Scene Albedo and Depth from a Single Foggy Image* Louis Kratz and Ko Nishino. In Proc. of IEEE Twelfth International Conference on Computer Vision, October, 2009.
- *Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models* Louis Kratz and Ko Nishino. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, June, 2009.