University of Central Florida
STARS

Electronic Theses and Dissertations, 2004-2019

2011

Analysis Of Behaviors In Crowd Videos

Ramin Mehran
University of Central Florida

Part of the Electrical and Electronics Commons
Find similar works at: https://stars.library.ucf.edu/etd
University of Central Florida Libraries http://library.ucf.edu

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation
Mehran, Ramin, "Analysis Of Behaviors In Crowd Videos" (2011). *Electronic Theses and Dissertations, 2004-2019*. 1773.
https://stars.library.ucf.edu/etd/1773

Showcase of Text, Archives, Research & Scholarship

ANALYSIS OF BEHAVIORS IN CROWD VIDEOS

by

RAMIN MEHRAN
B.S. K.N. Toosi University of Technology
M.S. K.N. Toosi University of Technology

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2011

Major Professor: Mubarak Shah

© 2011 RAMIN MEHRAN

ABSTRACT

In this dissertation, we address the problem of discovery and representation of group activity of humans and objects in a variety of scenarios, commonly encountered in vision applications. The overarching goal is to devise a discriminative representation of human motion in social settings, which captures a wide variety of human activities observable in video sequences. Such motion emerges from the collective behavior of individuals and their interactions and is a significant source of information typically employed for applications such as event detection, behavior recognition, and activity recognition. We present new representations of human group motion for static cameras, and propose algorithms for their application to variety of problems.

We first propose a method to model and learn the scene activity of a crowd using *Social Force Model* for the first time in the computer vision community. We present a method to densely estimate the interaction forces between people in a crowd, observed by a static camera. Latent Dirichlet Allocation (LDA) is used to learn the model of the normal activities over extended periods of time. Randomly selected spatio-temporal volumes of interaction forces are used to learn the model of normal behavior of the scene. The model encodes the latent topics of social interaction forces in the scene for normal behaviors. We classify a short video sequence of n frames as normal or abnormal by using the learnt model. Once a sequence of frames is classified as an abnormal,

the regions of anomalies in the abnormal frames are localized using the magnitude of interaction forces.

The representation and estimation framework proposed above, however, has a few limitations. This algorithm proposes to use a global estimation of the interaction forces within the crowd. It, therefore, is incapable of identifying different groups of objects based on motion or behavior in the scene. Although the algorithm is capable of learning the normal behavior and detects the abnormality, but it is incapable of capturing the dynamics of different behaviors.

To overcome these limitations, we then propose a method based on the Lagrangian framework for fluid dynamics, by introducing a *streakline* representation of flow. Streaklines are traced in a fluid flow by injecting color material, such as smoke or dye, which is transported with the flow and used for visualization. In the context of computer vision, streaklines may be used in a similar way to transport information about a scene, and they are obtained by repeatedly initializing a fixed grid of particles at each frame, then moving both current and past particles using optical flow. Streaklines are the locus of points that connect particles which originated from the same initial position.

This approach is advantageous over the previous representations in two aspects: first, its rich representation captures the dynamics of the crowd and changes in space and time in the scene where the optical flow representation is not enough, and second, this model is capable of discovering groups of similar behavior within a crowd scene by performing motion segmentation. We propose a method to distinguish different group behaviors such as divergent/convergent motion and lanes using this framework. Finally, we introduce *flow potentials* as a discriminative feature to

recognize crowd behaviors in a scene. Results of extensive experiments are presented for multiple real life crowd sequences involving pedestrian and vehicular traffic.

The proposed method exploits optical flow as the low level feature and performs integration and clustering to obtain coherent group motion patterns. However, we observe that in crowd video sequences, as well as a variety of other vision applications, the co-occurrence and inter-relation of motion patterns are the main characteristics of group behaviors. In other words, the group behavior of objects is a mixture of individual actions or behaviors in specific geometrical layout and temporal order.

We, therefore, propose a new representation for group behaviors of humans using the inter-relation of motion patterns in a scene. The representation is based on bag of visual phrases of spatio-temporal visual words. We present a method to match the high-order spatial layout of visual words that preserve the geometry of the visual words under similarity transformations. To perform the experiments we collected a dataset of group choreography performances from the YouTube website. The dataset currently contains four categories of group dances.

To my beloved wife,

my inspiration for life.

∼

To my lovely parents,

To my dear brother.

ACKNOWLEDGMENTS

I would like to thank Dr. Shah for his guidance and support. I am grateful to Mikel Rodriguez, Arslan Basharat, Vladimir Reilly, Paul Scovanner, Enrique Ortiz, Kishore Reddy, Berkan Solmaz, and Subhabrata Bhattacharya for their advice and inspiration. I am also grateful to Dr. Brian E. Moore, Dr. Gita Sukthankar, Dr. Marshall Tappen, and Dr. Aman Behal for serving on my committee.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Visual perception of objects, activities, and events are among the marvelous capabilities of the human mind that are developed early childhood. The human vision system is capable of performing computationally complicated tasks such as detecting or counting similar objects in a scene, in spite of occlusion and clutter, seemingly effortlessly. Research scientists in the computer vision community have been developing mathematical tools to detect objects, recognize objects and actions, and discover behaviors and events in visual scenes comparable to human capabilities. In all these efforts, the understanding of human activities is of a special interest for both application and research purposes. It paves the way for understanding human visual cognition and interaction skills. In addition, the scientific know-how is useful in a variety of applications such as surveillance and human-computer interaction.

Human activities are commonly social, and people perform actions in groups. In large groups of people, a crowd, people cooperate in different levels to achieve their goals. Different manifestations emerge from different levels of cooperation. At the highest level, people in a crowd cooperate as single entity toward a common goal such as people walking toward the exit of a stadium. At the second level, the cooperation of people in the crowd has local manifestations in achieving personal and group objectives in a scene of multiple entities. A traffic scene is a proper example of such a manifestation. Finally, at the most detailed level, people cooperate in inter-related groups to deliver

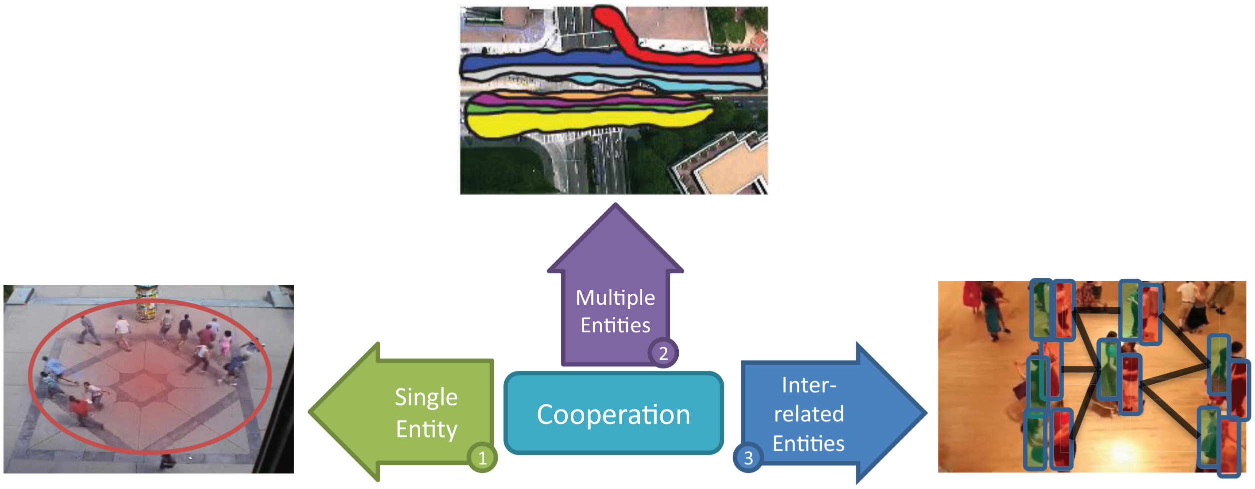


Figure 1.1: The cooperation of objects in a crowd scene has different manifestation: single entity, multiple but separate entities, and multiple and inter-related entities.

the goals of their group and achieve a common goal along the other entities. Group dances or team sports are examples of the last level of manifestations. Figure 1.1 summarizes the three levels of manifestations of cooperation in a crowd scene. The activities emerging from any of these manifestations are affiliated with complexities, such as emergent behaviors or self-organizing, which emerge from human interactions. These complexities per se make the visual understanding of crowd scene involving event detection, behavior recognition, group identification, and tracking a challenging problem. According to our experience and observations, in crowded scenes, even humans would face considerable challenges in performing simple visual recognition task such as counting or tracking objects. In part, these challenges arise because any increase in the number of objects in a scene raises the number of cases for visual inspection and recognition. More people in a scene would, furthermore, amplify the likelihood of occlusion and add to the clutter. This hinders the parallel perception and integration pathways in the human brain from exhibiting the *pop-up effect* to pinpoint the desired object [63]. Because of the challenges in visual recognition of crowded scenes, the traditional pipeline of object detection, tracking, and behavior analysis in computer vision research would fail poorly. Despite all of advances in computer vision research, human activity recognition and behavior analysis in crowds have remained as open problems due to the inherent complexity and vast diversity found in crowded scenes.

Crowd behavior analysis in computer vision research delivers new application domains, such as automatic detection of riots or chaotic acts in crowds, localization of the abnormal regions in scenes for high resolution analysis, group behavior recognition, and performance evaluation. This thesis develops algorithms and representations that provide an effective understanding of human activities in crowds by studying human interactions and group behaviors.

1.1 Overview and Motivation

Large political rallies, ceremonies, and rituals which involve large groups of people pose significant challenges for officials in terms of security, monitoring, and organization. First and foremost, in large gatherings the security of the event is of the highest importance. In dense crowds, any abnormal behavior or incident could lead to a cascade of undesirable events because of the synergic effect of human interactions [28]. Moreover, the larger the scale of the crowd becomes, the harder the visual surveillance is for the human eye. Therefore, authorities advocate the use of automatic tools to analyze crowd behavior and to assist the detection and localization of abnormal events within crowds.

In particular, certain behaviors in a crowd such as sudden convergence and divergence of the crowd flow are vitally important for forensic purposes and for the prevention of hazardous accidents in real time. For example, the annual Muslim Hajj in Mecca, Saudi Arabia, which is attended by millions of pilgrims, has increasingly suffered from stampedes, even as authorities have constructed new walkways and instituted other traffic controls to prevent them. Similar incidents have reported in India during Hindu religious holidays [3]. Moreover, stampedes may happen in social and political gatherings in case of crowd panic. For example, on 4th of March 2010, crowd panic emerged after hearing someone yelling the word ”bomb” and it created a stampede which caused injuring several people [1].

To improve the coordination of the crowds and to facilitate the flow of the people in public spaces, the transportation researchers are increasingly interested in ameliorating urban designs to adapt them to public needs and habits. For instance, by monitoring the behavior of the vehicular traffic in a highway and detecting the location of dense lanes and bottlenecks authorities efficiently manage the highway redesigns. Since most of the urban spaces are monitored by a network of cameras, computer vision algorithms can deliver invaluable support towards enhancing city design.

Another driving force in pursuing visual understanding of crowd behaviors is the limited number of automatic methods to recognize the collective behavior of people and their group interactions. In other words, the main body of work in surveillance and human action recognition focuses on individuals or small numbers of people. However, these methods are neither capable of nor designed for understanding and analysis of group behaviors such as panic, aggressive locomotion, lane formation, or group strategies in sports such as football or basketball. Not surprisingly, the vast majority of the literature has ignored the coordination of human actions and the emergent behaviors from these actions mainly because of the complexity of the action recognition for a single individual. Very limited research efforts have been done on understanding the coordinated human actions in groups, which itself can create a new area of research in human computer interaction. For instance, an automatic system to recognize and evaluate group performances or group sports would help instructors and learners improve the group executions tremendously. Moreover, such a method can be integrated with playing consoles such as Xbox, which is already equipped with an effective vision system, to access a large number of consumers. This thesis is an early step toward designing such systems for real-world application.

1.2 Nomenclature

In this section we define the terms we have employed to describe phenomenon related to crowded visual scenes:

- *Crowd* - corresponds to a group of objects in cooperation to achieve one or multiple shared goals. Depending on the number of people in a scene, it is divided into moderate, high, and extremely high density crowds. A crowd scene is a video stream that contains a crowd.

- *Crowd Behavior* - corresponds to the collective actions performed by the members of a crowd.

- *Activity* - corresponds to series of behaviors in a crowd scene which is dominantly performed to achieve a common goal.

- *Event* - corresponds to a sudden change in behaviors in a crowd scene.
- *Abnormal Event* and *Abnormal Behavior* - are used interchangeably, referring to an occurrence a behavior in time or space which is different from its learnt patterns.
- *Motion Segmentation* - corresponds to the task of dividing a crowd scene into spatio-temporal regions of distinct coherent motion.
- *Occlusion* - refers to inter-object occlusion resulting from the interactions of objects among themselves, and intra-object occlusion resulting from the interactions of objects with the scene.

1.3 Contributions

In this thesis, we present three major contributions. First, we have introduced Social Force model to the computer vision community as a method to understand the interaction and dynamics of groups of people in a crowd environment. Second, we have presented a novel understanding of motion flow in crowded scenes by introducing the streakline representation, a concept from fluid mechanics, to computer vision applications. The new representation not only enables us to model changes in the dynamics of the flow in the crowd, but it can benefit other applications in computer vision which use optical flow. We illustrate the ability of streaklines in flow segmentation and behavior recognition. Moreover, we acquire knowledge from fluid mechanics by investigating other representations of the flow, potential fields. We show that flow potentials are useful in understanding the underlying behavior of the crowd and to locate convergent and divergent regions and lanes. Finally, we propose an algorithm to model the geometrical relationships of spatio-temporal features in videos of groups of people to recognize complex behaviors that are composed of several coordinated motions.

In the following, we introduce our contributions in abnormal event detection, flow representation, and group behavior recognition in detail.

1.3.1 Abnormal behavior detection using Social Force Model

The first algorithm we develop in this thesis introduces a novel method to model the crowd behaviors. We observe that the behavior of the people in a scene is influenced by factors such as personal intentions, group coherency, and interactions. We present an algorithm to detect abnormal behaviors within a crowd by modeling the interactions in a scene over a temporal window. The main hypothesis is that the optical flow in a crowd scene is the product of the underlying dynamics of the crowd, and by observation of optical flow we are able to estimate the characteristics of the crowd dynamics such as the interaction force between individuals. Therefore, with the aid of a crowd dynamic model borrowed from human transportation research, we quantitatively measure the interactions between members of the crowd from the optical flow observations.

The core concept of this algorithm is *Social Force Model* which models the behavior of an individual person inside a crowded environment and is presented in a famous article by Helbing [28]. Here we present this model, for the first time to the computer vision community, to perform crowd behavior recognition and event detection. In the most general form, Social Force Model models the dynamics of the pedestrian which has been influenced by her/his personal desire and goal, group constraints, and environment or obstacles. To avoid the typical problems of visual understanding of crowd scenes such as occlusion and clutter, our method does not rely on tracking individuals or segmenting the scene. Instead, we propose a method to densely estimate the interaction forces between individuals in a crowd, using dense optical flow and particle advection in a Lagrangian framework. In this setup, we use videos of crowds which are obtained by a static camera. We propose a method to employ this model for understanding the behavior of a crowd in a holistic fashion.

To model the normal behavior of pedestrians in a scene, we observe the estimated interaction force over a window of time. The primary conjecture of this approach is that within a crowd with normal behavior, the interaction forces between individuals follow certain patterns that are distinct from unusual or abnormal behaviors. In this algorithm, the pattern of activities of the people in the scene is modeled in the form of spatio-temporal volumes of interaction forces. We use Latent Dirichlet Allocation (LDA) [11] to learn the model of the normal activities in a static camera over an extended period of time. By entering the set of interaction forces of normal video clips to LDA, it discovers the latent topics of social interaction forces in the scene for normal behaviors using a generative statistical model. The interaction forces are randomly selected as spatio-temporal volumes of data. In order to detect an abnormal behavior in the scene, we use a short sequence of n frames and estimate the interaction forces, and then we check the likelihood of normal behavior in the sequence using the learned model. If a sequence is classified as abnormal, the regions of anomalies in the abnormal frames are expected to appear in areas of high interaction. We localize these regions by thresholding the scalar field that contains the magnitudes of interaction forces.

1.3.2 Crowd flow segmentation and behavior recognition

The estimation of interaction forces within members of a crowd in the holistic framework, however, has certain limitations. The proposed algorithm uses a global estimation of the interaction forces within the crowd and disregards any explicit definition of groups in the crowd motion. It therefore, does not distinguish different groups of objects based on motion or behavior. For example, the method can estimate the interaction force between two groups of people that are walking toward each other, but it does not provide any information about the groups themselves and does not identify the behavior of people within the group. In addition, the algorithm proposes a model to learn the normal behavior and detects the abnormality by merely discovering a change in behaviors. However, the learned model is not capable of capturing the changes in the modes of behavior in dynamic scenes such as traffic lights with vehicular and pedestrian motion. Hence, in the second section of this thesis, we propose a method to overcome these limitations by performing flow segmentation to partition a crowd flow into groups of coherent behaviors and modeling the behaviors within each group. We introduce two new concepts from fluid dynamics to perform this task.

First, we introduce a *streakline* representation of flow based on dense optical flow and we juxtapose it against dense optical flow (i.e., *streamlines*), and dense particle trajectories (i.e., *pathlines*). Streaklines are traced in a fluid flow by injecting color material, such as smoke or dye, which is transported with the flow and used for visualization. In the context of computer vision, streaklines may be used in a similar way to transport information about a scene. Streaklines are the locus of points that connect particles which originated from the same initial position. We show that streaklines are useful in modeling the changes in the behavior of a crowd by capturing the spatio-temporal changes in the motion. We use streaklines to perform flow segmentation in dynamic scenes where the motion of the crowd switches between different modes of consistent and coherent motions such as scenes of traffic light in big cities.

Second, we introduce flow potentials as a tool for modeling crowd behaviors. In simplified mathematical models of fluids, it is often assumed that the fluid is incompressible and irrotational. These assumptions imply several conservation properties of the fluid, but most importantly, they lead to potential functions, which are scalar functions that characterize the flow in a unique way. For this discourse, potential functions enable accurate classification of behaviors in a scene, which is not possible with streak flow alone. In this representation, the crowd flow is decomposed into two separate scalar potential fields: *stream function* (incompressible), *velocity potential* (irrotational). In a broad view, the stream function provides the information regarding the steady and non-divergent part of the flow, whereas the velocity potential contains information regarding the local changes in the non-curling motions. Moreover, to have a complete picture of the flow we need information from both potential functions. With this perspective, we illustrate the strength of potentials in discriminating lanes and divergent/convergent regions within a crowd. Finally, we use flow potentials as a discriminative feature to recognize crowd behaviors in escape panic scenarios.

1.3.3 Discovery of Patterns in Group Behaviors

The proposed methods exploit optical flow as the low level feature and perform integration and clustering to obtain coherent group motion patterns and create a model of the behaviors in a bag of words framework regardless of the spatial layout of the features. We observe that in crowd video sequences, as well as, a variety of other vision applications such as part based action recognition, event detection, and activity recognition, the co-occurrence and inter-relation of motion patterns constitute the group behaviors. A prudent examination of the group behaviors reveals that every behavior or activity is a mixture of individual actions or behaviors in certain geometrical layout and temporal order. In order to harness the complexity of the inter-related behaviors in crowded scenes, we propose a behavior recognition algorithm that is focused on the spatial relation of behaviors. To achieve the goals, we propose a novel algorithm based on the bag of visual phrases of spatio-temporal visual words. The proposed method benefits from high-order spatial matching of visual words that preserve the geometry of their layout under similarity transformations. In other words, the proposed representation of behaviors is a bag of phrases which are invariant under arbitrary similarity transforms. In addition, we collected a new dataset which focuses on group behaviors specifically and we are sharing it with the research community. To perform the experiments we collected a dataset of group choreography performances from YouTube website. The dataset contains eight categories of group dances.

1.4 Organization of the Thesis

The thesis is structured as follows: **Chapter 2** reviews existing literature, and focuses on crowd behavior models in the computer vision literature as well as transportation research. The previous works on the representation of motion in crowd scenes and group behavior recognition are briefly reviewed. **Chapter 3** covers the problem of crowd as a single entity manifestation (cf. Figure 1.1), and presents a framework to perform event detection in crowded scenes based on Social

Force Model. **Chapter 4** considers the multiple entity manifestation of cooperation in crowds and proposes two new representations of flow based on Lagrangian framework to model multiple behaviors in dynamic crowded scenes. **Chapter 5** covers the inter-relations of entities in a crowd scene by presenting a novel representation of group behaviors based on spatial relationships of spatio-temporal features, and presents the first dataset of group choreography. Finally, the thesis is concluded in **Chapter 6** with a summary of contributions and description of future work.

CHAPTER 2: LITERATURE REVIEW

Crowd modeling and group behavior of large numbers of people and vehicles has been the subject of research studies for decades. However, visual understanding and analysis of crowds has remained a daunting task, mainly because of the complexity and diversity of group behaviors. In this chapter, we provide the context of this research by covering the most relevant research literature. We present the prominent works in crowd behavior modeling in transportation research and the application of these models in behavior recognition and pedestrian and vehicle tracking. In addition, we present some of the works in motion representation related to crowd scenes and their applications. Finally, we target the literature on group behavior recognition and activity recognition, and advances in bag of visual word representation to capture geometrical relationships of interacting motions.

2.1 Models of Crowds

2.1.1 Crowd Models in Transportation

Crowd behavior analysis is thoroughly studied in the field of transportation and public safety where some well-established models have been developed for describing the individual and group behaviors in crowded scenes [30][32]. At the high level, there are three main approaches in modeling the crowds in this community. (1) Microscopic approach, which defines pedestrians' motivation in movement and treats crowd behaviors as a result of a self-organization process. Social Force Model by Helbing et al. in [30] is the best known example of this approach. (2) Macroscopic approach, which focuses mainly on goal-oriented crowds. In this approach, a set of group-habits is determined based on the goals and destinations of the scene. Pedestrians are then partitioned into different groups to follow the predetermined habits. Therefore, instead of determining the motion of individuals the group behaviors are modeled [32][10]. (3) Hybrid methods, which inherit from macroscopic models as well as microscopic ones [50].

Based on socio-psychological studies, Helbing et al. in [30] originally introduced Social Force model to investigate the pedestrian movement dynamics. The social force captures the effect of the neighboring pedestrians and the environment on the movement of individuals in the crowd. Later, Helbing published his popular [28] work in combining the collective model of social panic with social force model to create a generalized model. In this model, both psychological and physical effects are considered in formulating the behavior of the crowd.

2.1.2 Crowd Models in Computer Vision

Recently, the computer vision community has focused on crowd behavior analysis. In [74] a review of the latest research trends and approaches from different research communities is provided. There are two main approaches in solving the problem of understanding crowd behaviors. In the conventional approach, which we refer as the "object-based" methods, a crowd is considered as a collection of individuals [65, 46]. Therefore, to understand the crowd behavior it is necessary to perform segmentation or detect objects to analyze group behaviors [12]. This approach faces considerable complexity in detection of objects, tracking trajectories, and recognizing activities in dense crowds where the whole process is affected by occlusions. On the other hand, "holistic" approaches [7][5] consider the crowd as a global entity in analysis of medium to high density scenes. In related works by Avidan et al. in [54] and Chan and Vasconcelos in [13], instead of tracking individual objects, scene modeling techniques are used to capture features for the crowd behavior and car traffic respectively. These are top-down approaches which directly tackle the problem of dense occluded crowds in contrast to the object-based methods. In addition, there are some works that mix the bottom-up view of object-based methods with top-down methods such as Ali and Shah's [6] for tracking humans in very dense crowds.

Moreover, Social Force model and its derivatives [62] have gain considerable attention in the computer vision community to perform different visual perception tasks such as tracking [53], motion prediction [45], and behavior recognition [62] with both bottom-up or to-down view.

2.1.3 Crowd Behaviors and Computer Graphics

Meanwhile, crowd behavior analysis has been an active research topic in simulation and graphic fields where the main goal is to create realistic crowd motions. The real crowd motion exhibits complex behaviors like line forming [32], laminar and turbulent flow [29][71], arching and clogging at exits, jams around obstacles [30], and panic [28]. Exact simulation of a crowd using behavior modeling leads to design of proper public environments that minimize the possibility of the hazardous events. Furthermore, in the graphics community, accurate modeling of the crowd movements is used to create realistic special effects of crowds without the need for human actors[60][17][42][64].

2.2 Motion Representation for Crowds

Several methods based on optical flow have been presented in recent years to handle the hurdles in visual understanding of crowds robust to occlusion and clutter. In computer vision, optical flow is widely used to compute pixel wise instantaneous motion between consecutive frames, and numerous methods are reported to efficiently compute accurate optical flow. However, optical flow does not capture long-range temporal dependencies, since it is based on just two frames, and by itself does not represent spatial and temporal features of a flow that are useful for general applications.

2.2.1 Lagrangian Approach

Recently, based on the Lagrangian framework of fluid dynamics, a notion of *particle flow* was introduced in computer vision. Particle flow is computed by moving a grid of particles with the optical flow through numerical integration, providing trajectories that relate a particle's initial position to its position at a later time. Impressive results employing particle flow have been demonstrated on crowd segmentation [5]. However, in particle flow the spatial changes may be ignored, and it has significant time delays.

Streaklines are well known in flow visualization [68, 31] and fluid mechanics [39] as a tool for measurement and analysis of the flow. With regard to flow visualization, streaklines are defined as the traces of a colored material in the flow. To understand streaklines, consider a fluid flow with an ink dye injected at a particular point. If the ink is continuously injected, then a line will be traced out by the ink in the direction of the flow; this is a streakline. If the direction of flow changes, then the streaklines change accordingly.

Streaklines are new to computer vision research. In this context, streaklines may be obtained by repeatedly initializing a grid of particles and moving all particles according to the optical flow, in the spirit of a Lagrangian fluid flow. In other words, place a particle at point p, and move the particle one time step with the optical flow. In the next time step, the point p is initialized with a new particle, then both particles are moved with the flow. Repeating this process on some time interval T produces particle positions from which we obtain streaklines.

2.2.2 Motion in Medium to High Density Crowds

In video scene analysis, which is the scope of this thesis, some approaches consider the entire scene as a collection of objects, and methods for scene understanding often involve object trajectory clustering and human action recognition. Examples include the tracking methods of [35] for individuals and [46] for groups of pedestrians, and the more recent work of Pellegrini et al. [53] in tracking based on social force model. Yet, the domain of application for these methods is limited to low density scenes with medium to high pixel resolutions on objects. Our work is concerned with high density scenes and low object resolution.

In other approaches, motion and tracking are represented by a set of modalities such as salient feature points [12, 24], and spatio-temporal volumes [36]. This promotes occlusion handling while preserving local accuracy. In the related approaches, it is common to represent both crowds and individuals as a set of regions, group of feature points, or sparse flows. In [12], Brostow and Cipolla use low level feature tracking to detect individuals in a dense crowd. Seemann et al. [59] presented a generative model to detect pedestrians as a combination of occupancy distributions.

Other methods of scene understanding involve particle tracking, motion pattern recognition, and segmentation based on dense optical flow [6, 55]. These methods are popular due to the intrinsic ability of global approaches to handle occlusion. The framework provides insight to social/group behavior of humans in crowds, but individual tracking or action recognition is only possible through a top-down framework. Recent works of Ali and Shah [5] on crowd analysis, and [37, 7] on abnormal behavior detection fall into this category. In addition, the particle video method [57] of Sand and Teller has a potential application in crowded scenes as it was originally introduced to handle occlusions while providing dense motion information.

2.3 Group Behavior Recognition

Another important application in the field of computer vision, that essentially relies on modeling motion, is the representation, recognition, and classification of human actions and activities. Group behavior recognition is studied under the umbrella of activity recognition which is a broad and active area of research, and comprehensive reviews of the available methods can be found in [14, 4]. In this section, we restrict the discussion of group behavior methods to methods which use bag of visual word paradigm for activity recognition as it is the most relevant part of the literature to this thesis.

2.3.1 Bag of Visual Words

Computer vision research in activity and action recognition have been inspired by the bag of words (BoW) representation from text categorization research. The adapted idea is commonly referred to as bag of visual words (BoVW) model for visual recognition. The process has two major stages: (1) vocabulary construction, and (2) category model learning. A vocabulary is constructed usually with k-means algorithm [19, 58, 69], and the category model learning is performed by a classifier such as SVM.

2.3.2 Advantages of the Bag of Words Paradigm

One major benefit of the bag of visual words representation is that does not require a background subtraction or object tracking as its building blocks [67, 51]. The empowering characteristic stems from the nature of local feature detectors such as $3D$ Harris corners [40] or Gabor filters [19]. In addition, the popular descriptors for interest points are usually scale, rotation, and translation invariant. Because of these endowments, the visual words method is receiving increasing attention in action and activity recognition [40, 19, 44, 25, 69, 72, 43].

2.3.3 Geometrical Relationships of Features

However, the representation discards any geometrical relationships of local features and ignores the shape, and the spatial structure. This limits the discriminative power of the method as many motion patterns are defined by their shape or the spatial layout. Many works [18, 20, 41, 49] have been done to model the spatial information by the locations of the local features or parts with respect to the activity or the object center. These works usually require a search for the object or activity center, which requires a prior knowledge and a separate level of learning. Methods like Constellation model [21] create a global spatial representation through a probabilistic mode of mutual relationship of the parts, and therefore can be applied directly without the need for prior knowledge. However, these methods suffer from high computation costs.

In order to over come these limitations, researchers have investigated the relationships of collocations of visual words or so called *visual phrases*. The common practice in computation visual phrases (groups of visual words) is to focus on co-occurrences of the words in entire frame or in a local neighborhood [73, 70, 26]. Co-occurrences in the entire images fail to encode the spatial information between. On the other hand, considering only the local neighborhoods falls short of considering long range relationships . In [77, 75], the authors introduce an effective method based on Generalized Hough Transform to overcome the limitations in the bag of words such that it considers the spatial layouts of the words in both local and global formations. Their method models the bag of phrases such that it is invariant to translation and scale, and only partially to rotation. The method can identify a set of visual words in a certain geometrical layout even if they have gone through arbitrary translation or scale. The authors in [77, 76, 75] suggested, but not did implement, the possible extension of their method to add complete rotation and scale invariance properly. We will discuss the details of these shortcomings in **Chapter** 5.

CHAPTER 3: SOCIAL FORCE MODEL FOR ANOMALY DETECTION

In this chapter we introduce a novel method to detect and localize abnormal behaviors in crowd videos using Social Force model [30]. For this purpose, a grid of particles is placed over the image and it is advected with the space-time average of optical flow. By treating the moving particles as individuals, their interaction forces are estimated using social force model. The interaction force is then mapped into the image plane to obtain Force Flow for every pixel in every frame. Randomly selected spatio-temporal volumes of Force Flow are used to model the normal behavior of the crowd. We classify frames as normal and abnormal by using a bag of words approach. The regions of anomalies in the abnormal frames are localized using interaction forces. The experiments are conducted on a publicly available dataset from University of Minnesota for escape panic scenarios and a challenging dataset of crowd videos taken from the web. The experiments show that the proposed method captures the dynamics of the crowd behavior successfully. In addition, we have shown that the social force approach outperforms similar approaches based on pure optical flow.

3.1 Overview of the Method

Social force model [30] describes the behavior of the crowd as the result of interaction of individuals. Therefore, the abnormal crowd behavior is essentially an eccentric state of the crowd interactions. Since social force model in [30] emulates the crowd dynamics with a high degree of

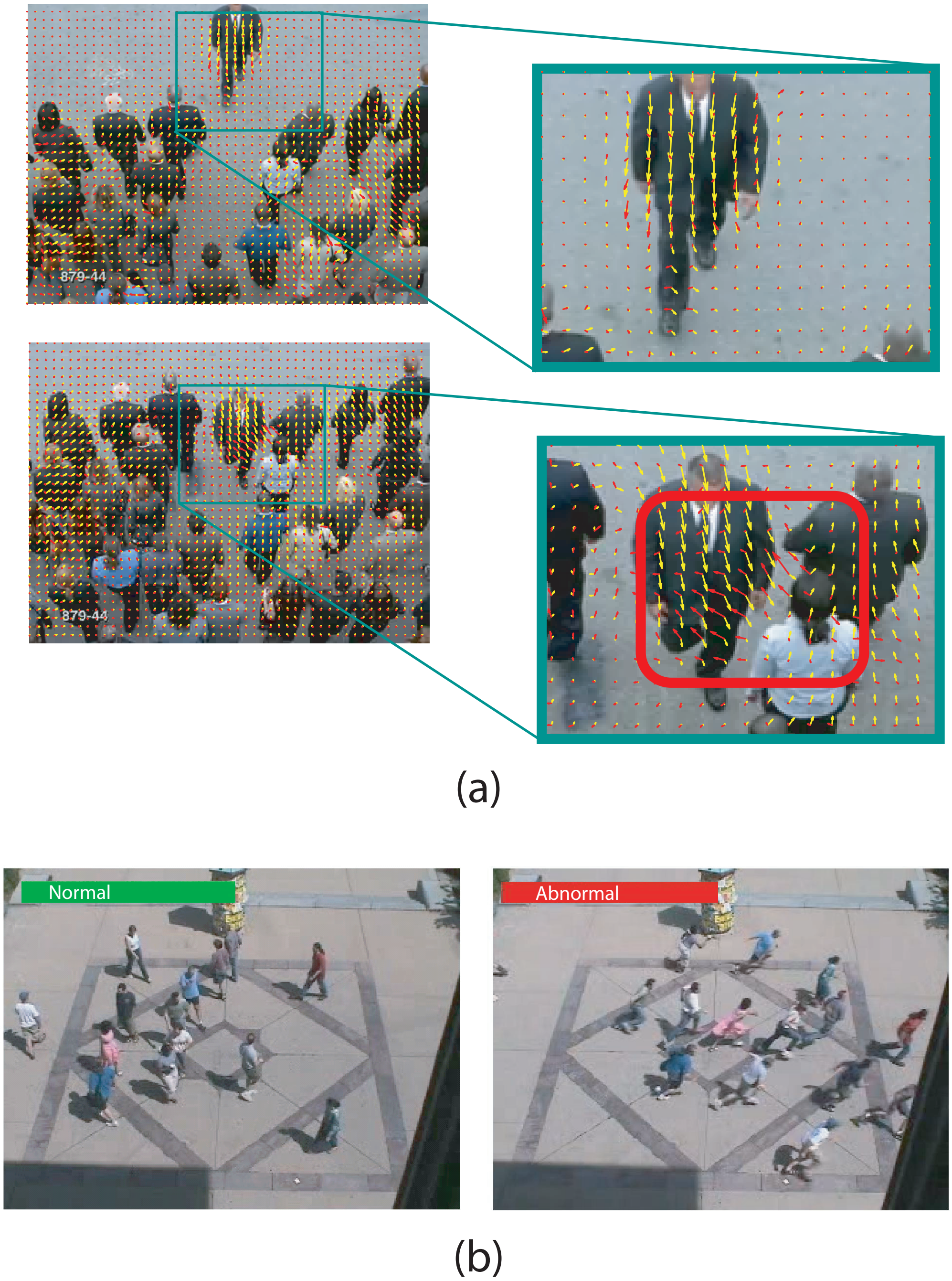


Figure 3.1: (a) The Optical flow (yellow) and the computed interaction force (red) vectors of two sampled frames. Note that the interaction force is computed accordingly for pedestrians who are approaching each other (red box). (b) An example of detection of escape panic using the proposed approach. Green denotes the normal and red denotes the abnormal frame.

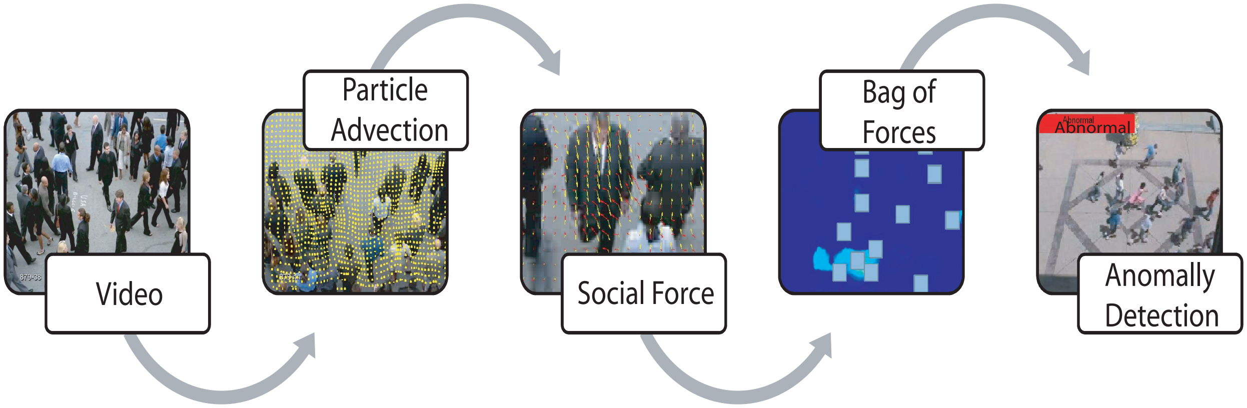


Figure 3.2: The summary of the proposed approach for abnormal behavior detection in the crowd videos.

accuracy, we conclude that abnormal social forces in the crowd portray abnormal behaviors. We estimate the social force parameters to create a model of likely behaviors in the crowd.

Figure 3.2 summarizes the main steps of the algorithm. In our method, we avoid tracking of objects to avert typical problems in tracking of high density crowds such as extensive clutter and dynamic occlusions. Instead, we incorporate a holistic approach to analyze videos of crowds using the particle advection method similar to [5]. In this approach, we place a grid of particles over the image and move them with the underlying flow field. We compute the social force between moving particles to extract interaction forces. In a crowd scene, the change of interaction forces in time determines the on going behavior of the crowd. We capture this by mapping the interaction forces to image frames. The resulting vector field is denoted as *force flow*, which is used to model the normal behaviors in a bag of words approach [11].

Andrade et al. [7] proposed a method for event detection in the crowd scene using Hidden Markov Model (HMM). However, the principal contribution of our work is to capture dynamics of the interaction forces in the crowd in addition to optical flow. Antonini et al. [9] reported a model for describing pedestrian behaviors to enhance tracking and detection. On the contrary, our primary goal is to introduce a holistic method independent of object tracking to detect abnormal crowd behaviors. Ali and Shah in [5] proposed a method for segmentation of high density crowds by introducing a method based on Coherent Structures from fluid dynamics and particle advection. Their method is capable of detecting instabilities in the crowd by identifying changes in the segmentation. Even though our work uses the same framework for particle advection, we use a completely different course by estimating the interaction forces of people in the crowd and detect anomalies directly without segmentation. In addition, the proposed method is not confined to high-density crowds in contrast to the method in Ali and Shah in [5].

The organization of this chapter is as follows. In the next section we introduce Social Force model for modeling the crowd movement. In Section 3.3 we introduce our method to estimate the social forces in the crowd. Section 3.4 describes the proposed method to detect abnormal behaviors in the crowd. Finally, in Section 3.5 we demonstrate abilities of the approach to detect and localize abnormal behaviors on a publicly available dataset.

3.2 Social Force Model

In the following, we describe social force model for pedestrian motion dynamics by considering personal motivations and environmental constraints. In this model, each of N pedestrians i with mass of m_i changes his/her velocity v_i as

$$m_i \frac{dv_i}{dt} = F_a = F_p + F_{int}, \tag{3.1}$$

as a result of actual force F_a, and due to individualistic goals or environmental constraints. This force consists of two main parts: (1) personal desire force F_p, and (2) interaction force F_{int}.

People in crowds generally seek certain goals and destinations in the environment. Thus, it is reasonable to consider each pedestrian to have a desired direction and velocity v_i^p. However, the crowd limits individual movement and the actual motion of pedestrian v_i would differ from the desired velocity. Furthermore, individuals tend to approach their desired velocity v_i^p based on the personal desire force

$$F_p = \frac{1}{\tau}(v_i^p - v_i), \tag{3.2}$$

where τ is the relaxation parameter.

The interaction force F_{int} consists of the repulsive and attraction force F_{ped} based on psychological tendency to keep a social distance between pedestrians and an environment force F_w to avoid hitting walls, buildings, and other obstacles. Therefore, the *interaction force* is defined as

$$F_{int} = F_{ped} + F_w. \tag{3.3}$$

It is logical to model pedestrians such that they keep small distances with people they are related or attracted to and keep far distances from discomforting individuals or environments. In social force model, these forces are defined based on potential fields functions. Further elaboration of this issue is not in the interest of this thesis and readers are referred to [30] and [28] for detailed discussion of these functions. In this chapter, we focus our attention to estimate the *interaction force* F_{int} between pedestrians as a single quantity.

Generalized social force model considers the effect of *panic* where herding behaviors appear in events like escaping from a hazardous incident. In this model, personal desire velocity v_i^p is replaced with

$$v_i^q = (1 - p_i)v_i^p + p_i\langle v_i^c\rangle, \tag{3.4}$$

where p_i is the panic weight parameter and $\langle v_i^c\rangle$ is the average velocity of the neighboring pedestrians. The pedestrian i exhibits individualistic behaviors as $p_i \rightarrow 0$ and herding behaviors as $p_i \rightarrow 1$. Overall, generalized social force model can be summarized as

$$m_i \frac{dv_i}{dt} = F_a = \frac{1}{\tau}(v_i^q - v_i) + F_{int}. \tag{3.5}$$

Generalized social force model is the cornerstone for many studies in simulation of crowd behavior [29] [71][38] in addition to the studies in computer graphics [52][61][60] for creating realistic animations of a crowd. Furthermore, estimation of parameters of the model provides valuable information about the governing dynamics of a crowd [34].

3.3 Estimation of Interaction Forces in Crowds

In this section, we describe the process of estimation of interaction forces F_{int} from a video of a crowd using social force model. The ideal case for computing the social force is to track all objects in the crowd and estimate the parameters as in [34]. However, tracking of individuals in a high density crowd is still a challenging problem in computer vision [6]. In a nutshell, low resolution images of the objects in the dense crowd, dynamic and static occlusions, and similarity of the objects have made the tracking of individuals in the crowd a daunting task. Therefore, in the crowded scenes, object-based methods fall short in accurate estimation of social force parameters.

It has been observed that when people are densely packed, individual movement is restricted and members of the crowd can be considered granular particles [6]. Thus, in the process of estimating the interaction forces, we treat the crowd as a collection of interacting particles. Similar to [5], we put a grid of particles over the image frame and move them with the flow field computed from the optical flow. To analyze the scene, we treat moving particles as the main cue instead of tracking individual objects. As the outcome, the proposed method does not depend on tracking of objects; therefore, it is effective for the high density crowd scenes as well as low density scenes. Furthermore, the particle advection captures the continuity of the crowd flow which neither optical flow nor any instantaneous measure could capture [56] [5].

In the next section we describe a modification of social force model to operate on moving particles instead of pedestrians and we discuss the advection of particles using the optical flow. In Section 3.3.2, we introduce the modification of the generalized social force model for particle advection.

3.3.1 Particle Advection

To advect particles, we compute the average optical flow field O_{ave}, which is the average of the optical flow over a fixed window of time and as well as space. The spatial average is done by a weighted average using a gaussian kernel. To start the particle advection process, we put a grid of

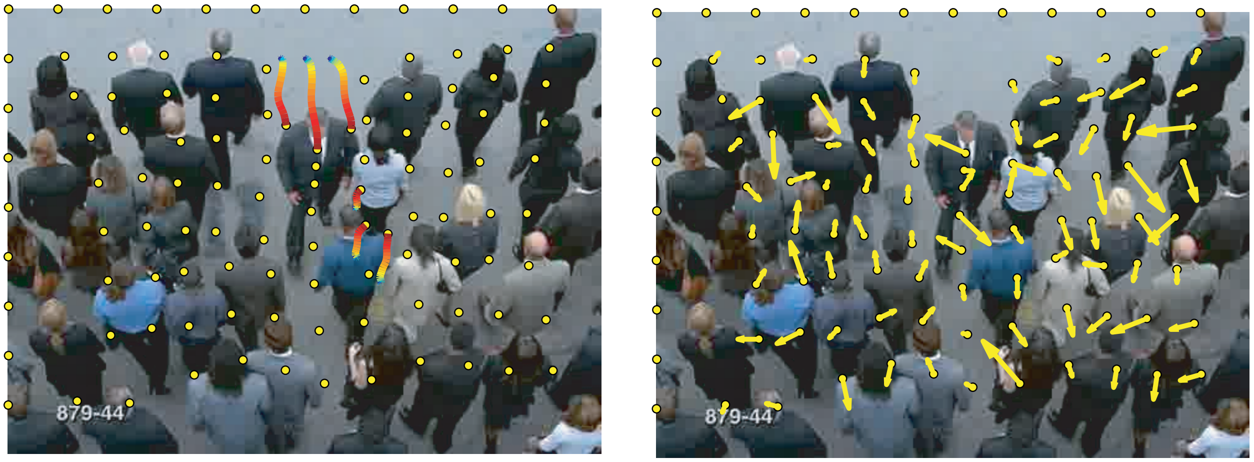


Figure 3.3: An example of particle advection using the average optical flow field and the corresponding interaction forces. (Left) The trajectories of a small set of particles are depicted for demonstration. (Right) The set of computed interaction forces of particles.

N particles over the image and move the particles with the corresponding flow field. The effective velocity of particles is computed using a bilinear interpolation of the neighboring flow field vectors.

Using the described particle advection process, particles move with the average velocity of their neighbors. This resembles the collective velocity of a group of people in the crowd. Figure 3.3 illustrates a example of particle advection.

3.3.2 Computing the Social Force

As a tangible analogy, the particles moving by optical flow resemble the motion of the leaves over a flow of water. This notion helps in understanding the modification of social force model for the particle grid. In the case of leaves, wherever there is an obstacle, joining, or branching of the fluid, the leaves have different velocities than the average flow. By analogy, we conclude that particles are also capable of revealing divergent flows in the regions that their desired movement is different from the average flow.

We modify Equation 3.5 for particle advection by defining the actual velocity of the particle v_i as

$$v_i = O_{ave}(x_i, y_i), \tag{3.6}$$

where $O_{ave}(x_i, y_i)$ is the effective spatio-temporal average of optical flow for the particle i with the coordinate (x_i, y_i). We write the desired velocity of the particle v_i^q as

$$v_i^q = (1 - p_i)O(x_i, y_i) + p_i O_{ave}(x_i, y_i), \tag{3.7}$$

where $O(x_i, y_i)$ is the optical flow of particle i with coordinate (x_i, y_i). The effective average flow field and effective optical flow of particles are computed using linear interpolation.

Using the above modification, particles move with the collective velocity of the flow of the crowd. Furthermore, each particle has a desired velocity which depends on the current optical flow. Hence, any difference between the desired velocity of the particle and its actual velocity relates to interaction of the particle with the neighboring particles or the environment. Figure 3.3 demonstrates an example of the computed interaction force for a sub-sample set of particles.

Without loss of generality, for a given scene or certain type of crowd with consistently similar sizes of objects, we assume that $m_i = 1$. Hence, we can simply estimate interaction force, F_{int}, from equation 3.5 for every particle as

$$F_{int} = \frac{1}{\tau}(v_i^q - v_i) - \frac{dv_i}{dt}, \tag{3.8}$$

where $\frac{dv_i}{dt}$ is computed with a difference approximation.

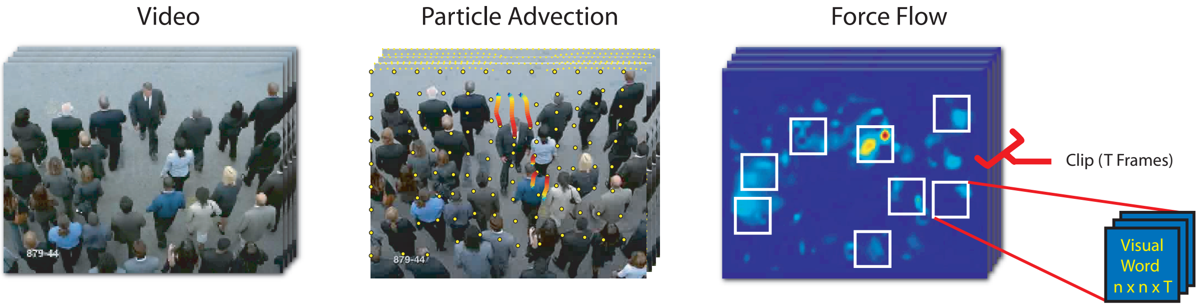


Figure 3.4: The overall demonstration of the algorithm. Using the average optical flow field, a grid of particles is updated and the interaction forces between particles are computed. The forces are mapped back to the image space to construct the force flow. Visual words are randomly picked as 3D volumes of features from the force flow to use in LDA model.

3.4 Event Detection

The computed interaction forces determine the synergy between advecting particles. However, discrete value of forces is not a clear evidence of abnormal behaviors. For instance, in a normal scene of a stock market, the interaction force of stock brokers would be quite higher than the interaction forces of walking pedestrians in a street scene. In other words, the instantaneous forces in a scene do not discriminate the abnormalities but the pattern of forces over a period of time does. In the following, we propose a method to model the normal patterns of forces over time.

In this method, we map the magnitude of the interaction force vectors to the image plane such that for every pixel in the frame there is a corresponding force vector. As a result, for a stream of image frames $I(t)$ of m pixels, we construct a feature matrix of *force flow* $S_f(t)$ of the same resolution. Figure 3.5 illustrates force flow for a sample of frames of a video stream.

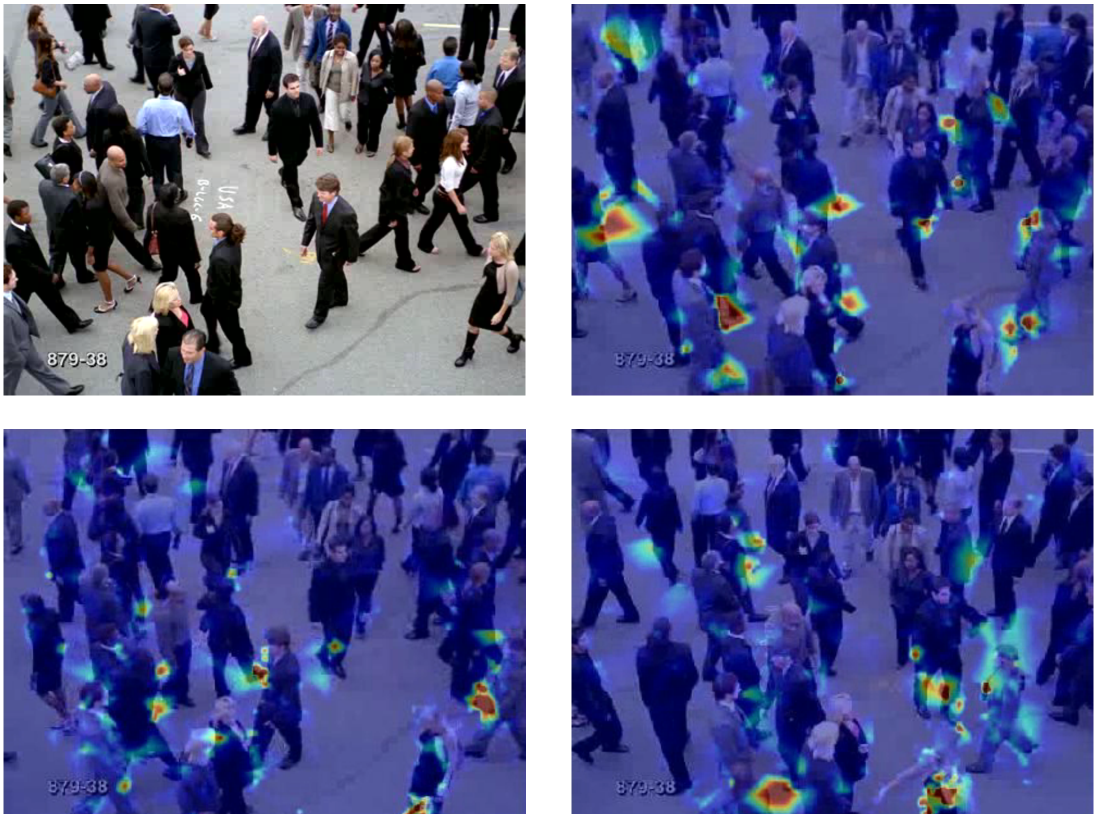


Figure 3.5: Examples of the computed force field for one example video sequence. The image on the top left is the first frame, and the rest are sample frames of the sequence with alpha channel of forces overlayed. The color map Jet is used so red values represent higher forces where as blue values represent low force flow. The high values of forces are results of two cases: (1) a large difference between instantaneous optical flow and the spatio-temporal average of optical flow or (2) a large change in spatio-temporal average of optical flow between two frames.

The process of identifying the likely patterns in the $S_f(t)$ is a special case of scene modeling which is considerably studied in computer vision. The bag of words [11] method is one of the typical candidates for such an analysis. In this chapter, we consider using bag of words method to estimate the likelihood force flow $S_f(t)$ and we use only normal videos for training LDA.

To use LDA, we partition the force flow into blocks of T frames which we refer to as *Clips*. Next, from each clip D_j, K *visual words* Z_j are extracted. We randomly pick visual words of size $n \times n \times T$ from locations in the *force flow* where corresponding optical flow is not zero. Finally, a code book of size C is formed using K-means clustering. Figure 3.4 illustrates the process of computing *force flow* and the extraction of visual words.

Therefore, for a set of normal force flows of a given scene or a group of similar scenes, we construct the corpus $D = \{D_1, D_2, D_3, ..., D_M\}$ and we use Latent Dirichlet Allocation (LDA) [11] to discover the distribution of L topics for the normal crowd behavior. Using the provided code for modified Expectation Maximization (EM) algorithm in [11], we approximate the bag of words model to maximize the likelihood of corpus as

$$\ell(\alpha, \beta) = \sum_{j=1}^{M} \log p(D_j|\alpha, \beta), \tag{3.9}$$

where α is the first learnt parameter, and it defines the Dirichlet distribution that generates the random variable to generate the latent topics. β is the second learnt parameter that is endowed with a posterior distribution that influences the generation of words in cooperation with the latent topics. By using the model, we estimate the likelihood $\log p(D_j|\alpha, \beta)$ for every clip from the video sequence. Based on a fixed threshold on the estimated likelihood, we label frames as normal or as abnormal.

3.4.1 Localization of Abnormalities

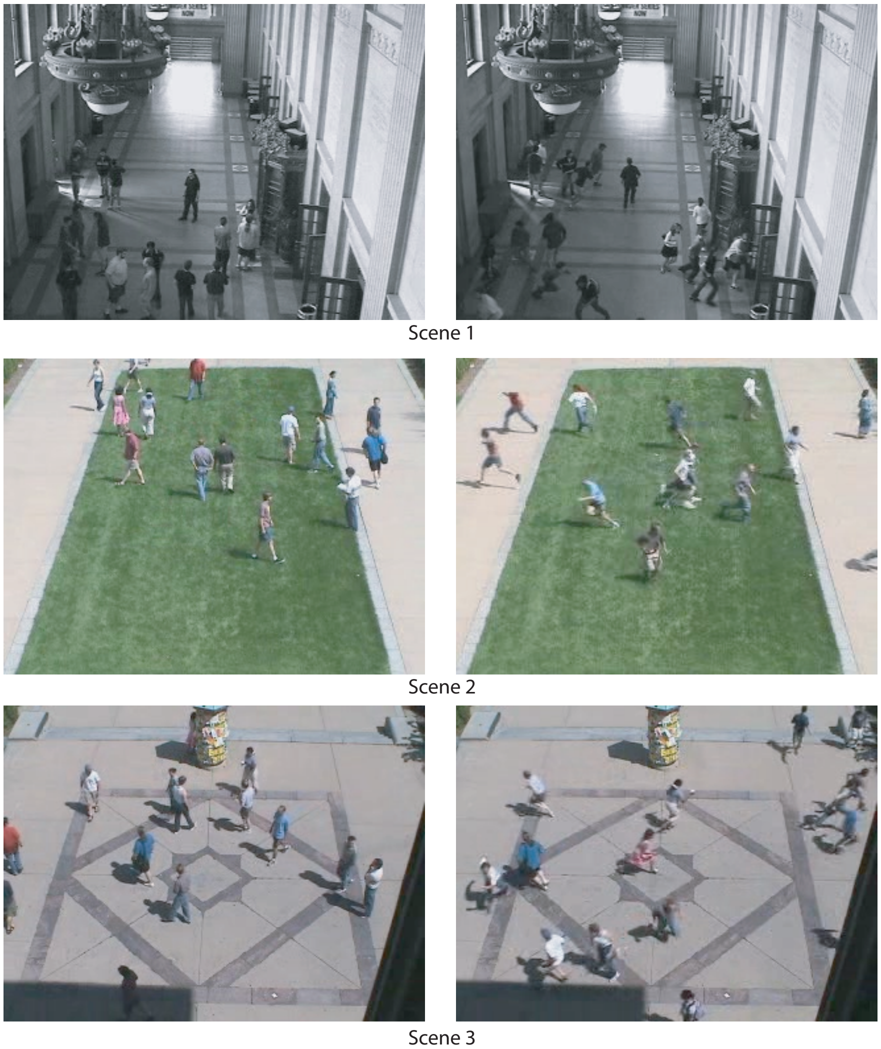
Using LDA model with force flows, we distinguish abnormal frames from the normal frames. Although it is helpful to localize regions in the frame that correspond to the abnormalities, the bag of words method does not implicitly provide a method to localize the unlikely visual words. As we discussed earlier, the force flow reveals the interaction forces in the scene, which correspond to the activities in the scene. In an abnormal scene, we expect the anomalies to occur in active regions or the regions with higher social interactions. Therefore, we localize abnormalities in the abnormal frame by locating the regions of high force flow.

3.5 Experiments and Discussion

3.5.1 The UMN Dataset

The approach is tested on the publicly available dataset of normal and abnormal crowd videos from University of Minnesota [2]. The dataset comprises the videos of 11 different scenarios of an escape event in 3 different indoor and outdoor scenes. Figure 3.6 shows sample frames of these scenes. Each video consists of an initial part of normal behavior and ends with sequences of the abnormal behavior.

In the particle advection phase, the resolution of the particle grid is kept at 25% of the number of pixels in the flow field for computational simplicity. For computation of the interaction forces, the panic parameter is kept fixed as $p_i = 0$. Therefore, the interaction forces are computed by assuming that the crowd is not in panic in normal motion. As a result, any high magnitude interaction


Scene 1

Scene 2

Scene 3

Figure 3.6: Sample frames in three different scenes of the UMN dataset: Normal (left) and abnormal (right).

force relates to activities different from the collective movement of the crowd. The force flow is computed by linear mapping of the force field into an image of the same resolution as the video frame. For construction of visual words, we used 3D volumes of $5 \times 5 \times 10$. $K = 30$ visual words are extracted from blocks of $T = 10$ frames of force flow with one frame overlap. The final codebook contains $C = 10$ clips. The LDA is used to learn $L = 30$ latent topics.

To evaluate the approach, 5 different video sequences of the first scene are selected and LDA model is created for visual words from the frames with normal behavior. The trained model is used to estimate the likelihood of being normal for blocks of T frames. Therefore, the method chops any input video into clips of T frames and labels all frames in each clip as normal or abnormal. Figure 3.7 shows some of the qualitative results for detection of abnormal scenes. In each row, the figure depicts the first frame of the sequence on the left and a detected abnormal frame on the right. The black triangles on the horizontal bars identify the timing of the shown abnormal frames. The false positive detections in Figure 3.7 are the result of incorrect estimation of social forces. Overall, these results show that estimated social force model is capable of detecting the governing dynamics of the abnormal behavior, even in the scenes for which it is not trained. All videos in the dataset exhibit behavior of escape panic and the proposed approach successfully models the dynamics of the abnormal behavior regardless of the scene characteristics.

In addition, we demonstrate the power of the proposed social force model in capturing the abnormal behaviors in contrast to use of optical flow. In this experiment, instead of force flow, we use spatio-temporal patches of optical flow as visual words. Thus, we create a codebook from optical flow information to learn an LDA model. We use the same parameters for LDA training in the experiment with optical flow. Therefore, the blocks of 10 frames of the magnitude of the optical flow are used as clips to learn the distribution of latent topics and to compute the likelihood of frames. We use the same dataset for this experiment with the same set of parameters for

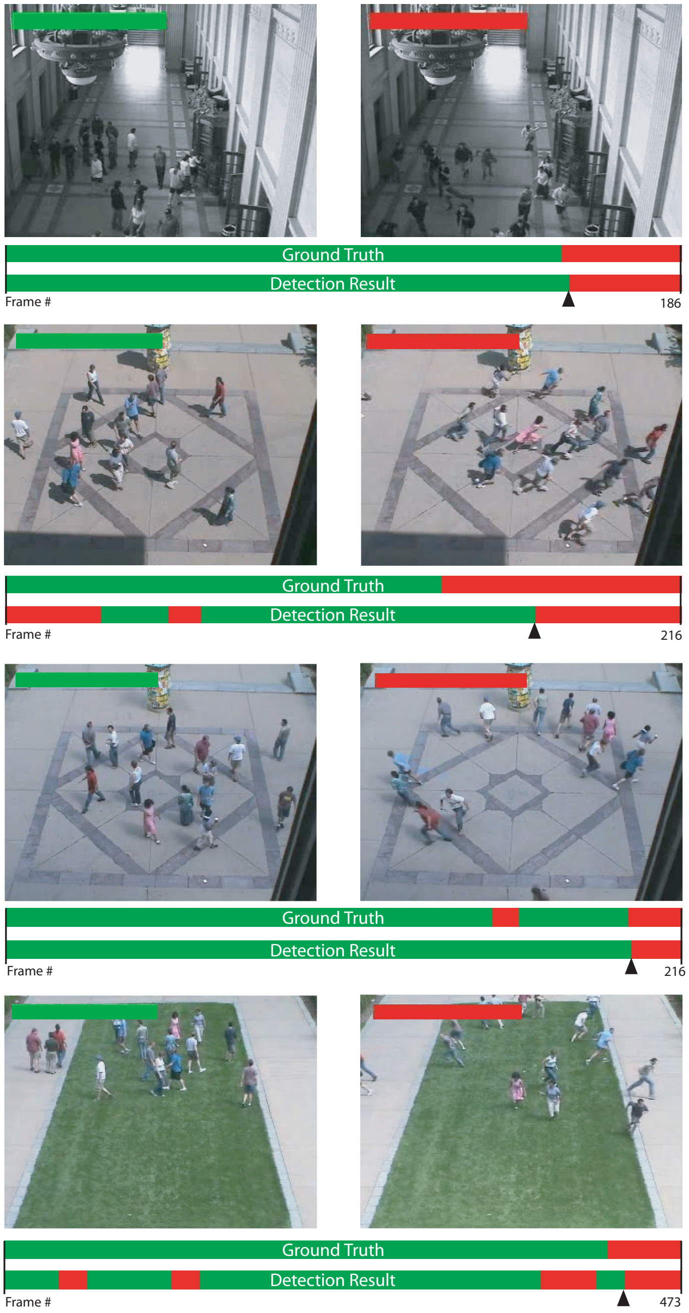


Figure 3.7: The qualitative results of the abnormal behavior detection for four sample videos of UMN dataset. Each row represents the results for a video in the dataset. The ground truth bar and the detection bar represent the labels of each frame for that video. Green color represents the normal frames and red corresponds to abnormal frames. The left column shows the first frame of the video and the right column is the first frame of the detected abnormal block (black triangles).

Method	Area under ROC
Social Force	0.96
Pure Optical Flow	0.84

Table 3.1: The comparison of the use of the proposed social force method and pure optical flow for detection of the abnormal behaviors in the UMN dataset.

learning LDA model. The ROC curves in Figure 3.9 illustrate that the proposed method outperforms the method based on pure optical flow in detecting abnormalities, and Table 3.1 provides the quantitative results of the comparison.

In Figure 3.8, we demonstrate the qualitative results of localization of abnormal behaviors in the crowd, where the escaping individuals are highlighted as abnormal areas of frames. The results show that the interaction forces are capable of locating the abnormalities in the regions that are occupied by the crowd. As the figure shows, the proposed method provides regions of abnormality and does not label individuals.

3.5.2 The Web Dataset

To evaluate our method in practical applications, we conduct an experiment on a challenging set of videos which has been collected from the sites like Getty Images and ThoughtEquity.com which contain documentary and high quality videos of crowds in different urban scenes. The dataset comprises 12 sequences of normal crowd scenes such as pedestrian walking, marathon running, and 8 scenes of escape panic, protesters clashing, and crowd fighting as abnormal scenes. All the

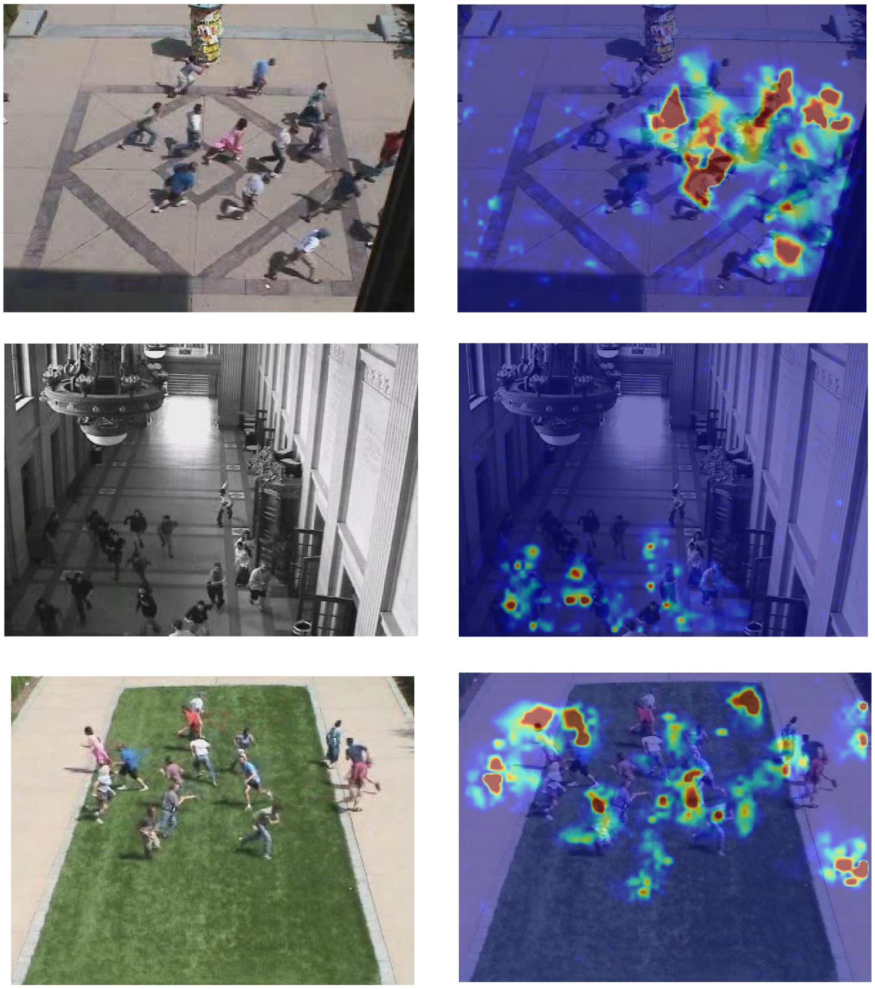


Figure 3.8: The localization of the abnormal behaviors in the frames using the interaction force. Original frames (left), Localized abnormal behaviors (right). Red pixels correspond to the highly probable abnormal regions.

frames are resized to the fixed width of 480 pixels. Figure 3.10 shows sample frames of the normal and abnormal sequences.

In this experiment, the resolution of the particle grid is kept at 10% of the number of original pixels. For construction of visual words, we extracted $K = 30$ similar $5 \times 5 \times 10$ volumes from

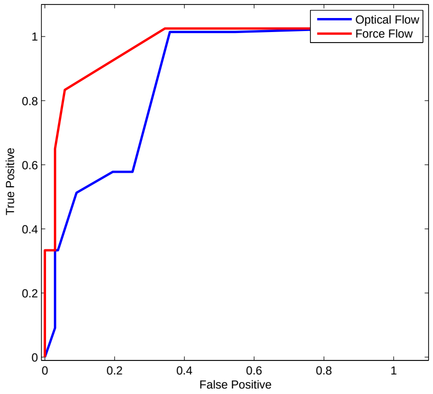


Figure 3.9: The ROCs for detection of abnormal frames in the UMN dataset. Proposed method (Red) outperforms use of pure optical flow (Blue).

a block of $T = 10$ frames of force flow. The codebook for this experiment contains $C = 30$ clips and the LDA is used to learn $L = 50$ latent topics. To learn the LDA model, we used the normal sequences in a 2-fold fashion. We randomly excluded 2 sequences from the normal set and trained on the rest. In the testing phase we added the excluded sequences to the test set. We did this experiment 10 times and constructed the ROC by averaging the results of these experiments.

The ROC in Figure 3.11 demonstrates that the proposed method outperforms optical flow method to distinguish abnormal sequences.

3.6 Conclusion

Using social force model, we introduce a method to detect abnormal behaviors in crowd scenes. We address the ability of the method to capture the dynamic of crowd behavior based on the interaction forces of individuals without the need to track objects individually or perform segmentation.

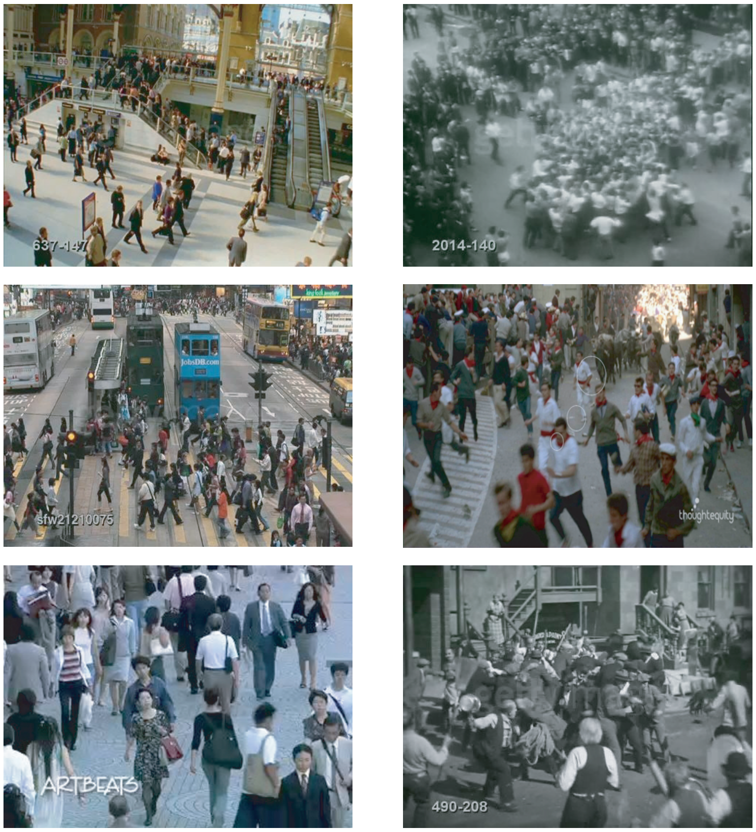


Figure 3.10: Sample frames of 6 sequences of our web dataset. (Left Column) Normal samples. (Right column) Abnormal samples.

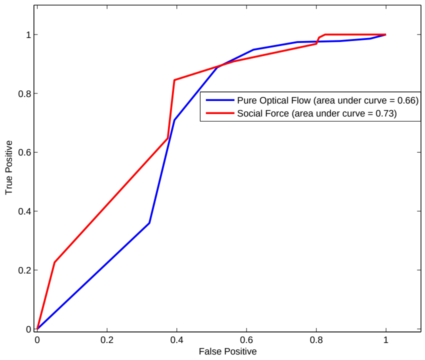


Figure 3.11: The ROCs of abnormal behavior detection in the web dataset.

The results of our method, indicates that the method is effective in detection and localization of abnormal behaviors in the crowd.

CHAPTER 4: STREAKLINES REPRESENTATION OF FLOW

Based on the Lagrangian framework for fluid dynamics, a streakline representation of flow is presented to solve computer vision problems involving crowd and traffic flow. Streaklines are traced in a fluid flow by injecting color material, such as smoke or dye, which is transported with the flow and used for visualization. In the context of computer vision, streaklines may be used in a similar way to transport information about a scene, and they are obtained by repeatedly initializing a fixed grid of particles at each frame, then moving both current and past particles using optical flow. Streaklines are the locus of points that connect particles which originated from the same initial position. In this chapter, a streakline technique is developed to compute several important aspects of a scene, such as flow and potential functions using the Helmholtz decomposition theorem. This leads to a representation of the flow that more accurately recognizes spatial and temporal changes in the scene, compared with other commonly used flow representations. Applications of the technique to segmentation and behavior analysis provide comparison to previously employed techniques, showing that the streakline method outperforms the state-of-the-art in segmentation, and opens a new domain of application for crowd analysis based on potentials.

4.1 Introduction

Behavior analysis in crowded scenes remains an open problem in computer vision due to the inherent complexity and vast diversity found in such scenes. One hurdle, that must be overcome, is finding good ways to identify flow patterns without tracking individual objects, which is both impractical and unnecessary in the context of dense crowds. Another hurdle is finding good ways to understand changes in behavior when the scene context and crowd dynamics can vary over such a wide range.

In this chapter, we offer three major contributions. *First*, we assert a streakline framework as a new tool for analysis of crowd videos. We demonstrate streaklines can be more informative than commonly used flow representations, known as optical flow and particle flow. *Second*, we present an innovative algorithm to compute a fluid-like flow of crowds to perform behavior analysis. *Third*, we present potential functions as valuable tools for behavior analysis and complementing the streakline framework.

The capabilities of the streakline framework are tested in two applications: crowd segmentation and abnormal behavior detection. The segmentation results demonstrate an improvement for unsteady flows in comparison to state of the art. The behavior detection results show an improvement over base-line optical flow.

Table 4.1: Advantages of Streaklines over Streamlines and Pathlines

Streamlines	Pathlines	Streaklines
Spatial gaps in flow.	Ignores spatial changes.	Fills gaps.
Rough transitions in time.	Time delays.	Captures instant changes.

4.2 Streaklines vs. Pathlines and Streamlines

In fluid mechanics there are different vector field representations of the flow [39]:

Streamlines *are tangent to the velocity vectors at every point in the flow.* These correspond to optical flow, and a visual example is given in Figure 4.1(a).

Pathlines *are trajectories that individual particles in a fluid flow will follow.* These directly correspond to integration of optical flow in time and are illustrated by a set of curves with the spectrum of colors from Blue to Orange in Figure 4.1(b). Particle flow is the set of pathlines which are computed from time averaged optical flow [5].

Streaklines *represent the locations of all particles at a given time that passed through a particular point.* Figure 4.1(c) shows streaklines as red curves next to pathlines.

For flows that are steady and unchanging, these three representations are the same, but for flows that are unsteady, so that directions of flow can change with time, they are different. Since we are using a Lagrangian model for fluid flow to exploit the dynamics in crowd videos, where frequent changes in the flow are expected, it is important to know which vector field representation is most appropriate for the given problem. In this work, we provide a juxtaposition of streaklines with

Table 4.2: A table of values for x-coordinate particle positions, which are computed from the optical flow. Columns correspond to pathlines and rows correspond to streaklines.

	$L^p(0,T)$	$L^p(1,T)$	$L^p(2,T)$	$\cdots$	$L^p(t,T)$	$\cdots$	$L^p(T,T)$
$S^p(0,0)$	$x_0^p(0)$						
$S^p(0,1)$	$x_0^p(1)$	$x_1^p(1)$					
$S^p(0,2)$	$x_0^p(2)$	$x_1^p(2)$	$x_2^p(2)$				
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$			
$S^p(0,t)$	$x_0^p(t)$	$x_1^p(t)$	$x_2^p(t)$	$\cdots$	$x_t^p(t)$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$	
$S^p(0,T)$	$x_0^p(T)$	$x_1^p(T)$	$x_2^p(T)$	$\cdots$	$x_t^p(T)$	$\cdots$	$x_T^p(T)$

streamlines and pathlines, which correspond to commonly used methods [7, 17] based on optical flow and particle flow, respectively. Our theory and results show that streamlines leave spatial gaps in the flow, as well as choppy transitions between frames. This is because it is produced from instantaneous velocity vectors. Hence, this approach does not produce fluid-like flow for crowd videos [33]. Pathlines overcome this problem by filling the spatial gaps, but do not allow for detection of local spatial changes, and in addition create an artificial time lag. Our streakline approach provides solutions to each of these problems, and Table 4.1 gives an overview of the advantages.

To explain how streaklines are computed, let $(x_i^p(t), y_i^p(t))$ be a particle position at time t, initialized at point p and frame i for $i, t = 0, 1, 2, \ldots, T$. Repeated initialization at p implies $(x_i^p(i), y_i^p(i)) = (x_0^p(0), y_0^p(0))$. Particle advection is achieved by

$$\begin{aligned} x_i^p(t+1) &= x_i^p(t) + u(x_i^p(t), y_i^p(t), t) \\ y_i^p(t+1) &= y_i^p(t) + v(x_i^p(t), y_i^p(t), t)\ , \end{aligned} \tag{4.1}$$

where u and v represent the velocity field obtained from optical flow. This yields a family of curves, all starting at point p and tracing the path of the flow from that point in frame i. Naturally,

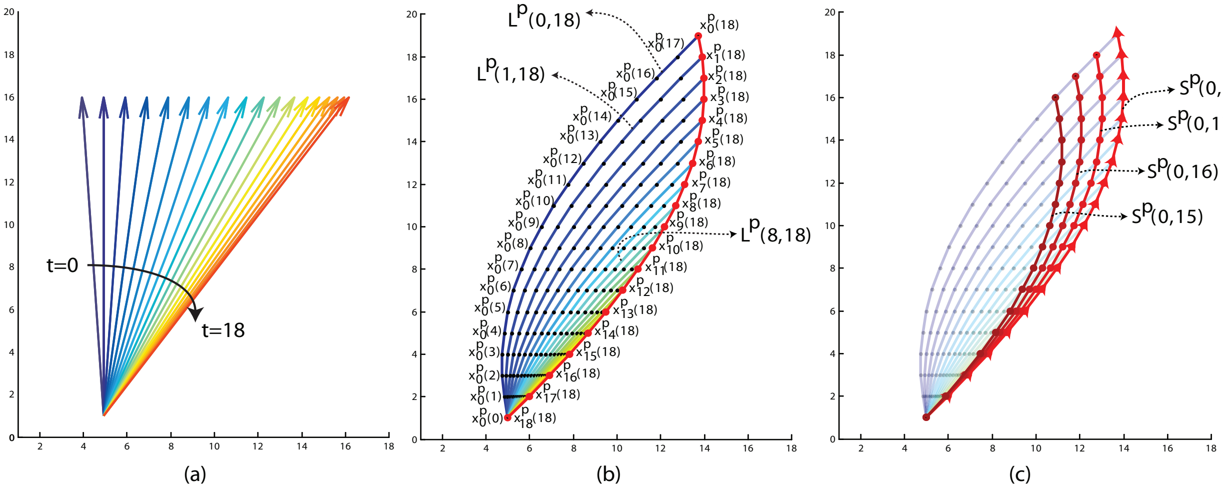


Figure 4.1: An illustration of pathlines and streaklines generated using a locally uniform flow field which changes over time. (Labels on points and curves directly correspond to Table 4.2.) (a) The changes in the flow vectors over time period $t = 0$ to $t = 18$. (b) The pathlines are illustrated as a spectrum of lines. Blue corresponds to the initiating frame of $t = 0$ and orange corresponds to initiating frame of $t = 18$. The red line illustrates the streakline at frame $t = 18$. (c) Streaklines at different frames as red curves to illustrate the evolution of the streaklines through time. The streakline at time $t = 18$ is illustrated along with the initiating motion vector as explained by (4.2).

for steady flow all these curves lie along the same path, but for unsteady flows the curves vary in direction and shape, characteristic of pedestrian flow.

Particle advection for all $i, t = 0, 1, 2, \ldots, T$ using (4.1), yields a table of values for $x_i^p(t)$ (shown in Table 4.2) and similarly for $y_i^p(t)$. The columns of the table show the pathlines $L^p(t, T)$, which are the particle trajectories from time t to T. The rows provide the streaklines $S^p(0, t)$, connecting all particles from t frames that originated at point p. Corresponding to this table, Figure 4.1 illustrates the set of streaklines and pathlines for an example unsteady flow at time $t = T$. At the start of observation, particles are initiated at every time instant at point p. The spectrum of lines from blue to orange represents the pathlines of particles which have been initiated at time $t = 0$. The solid red color lines depict streaklines. Since the flow is not steady, the streaklines and pathlines are different.

The unsteady flow at a point can be represented by either a set of pathlines or a streakline. However, the streakline provides a speed and memory gain, as a streakline with L particles corresponds to L pathlines with $L \times (L-1)/2$ particles. There are other interesting, less obvious, properties that streaklines inherit from fluid mechanics. First, in unsteady flows, extra long streaklines may exhibit shapes inconsistent with the actual flow, meaning they can not be allowed to get too long [27]. Second, as invented for visualization purposes, streaklines in fluids transport a color material along the flow, meaning they propagate changes in the flow along their path. Similarly, our setup allows streaklines to propagate velocities, given by the instantaneous optical flow $\Omega = (u, v)^T$ at the time of initialization, along the flow like a material. To this end, we define an *extended particle* i as a set of position and initial velocity

$$P_i = \{x_i(t), y_i(t), u_i, v_i\}, \tag{4.2}$$

where $u_i = u(x_i^p(i), y_i^p(i), i)$, and $v_i = v(x_i^p(i), y_i^p(i), i)$. In the whole scene, we consider only streaklines comprising extended particles. Figure 4.2 depicts streaklines for an example sequence.

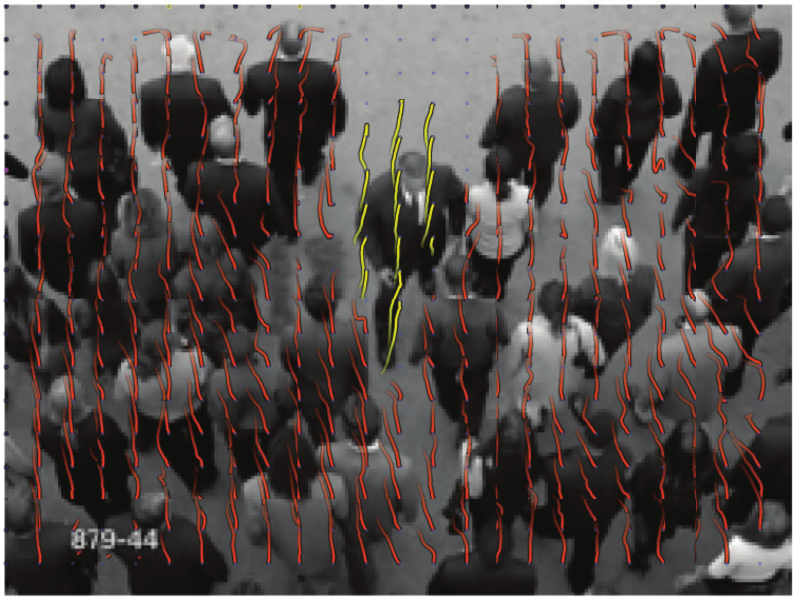


Figure 4.2: An illustration of streaklines for a video sequence.

4.3 Computations with Streaklines

Streaklines provide a means to recognize spatial and temporal changes in the flow, that neither streamlines nor pathlines could provide directly. This point is made here using streak flow and potential functions. In essence, streak flow is obtained by time integration of the velocity field, while potential functions are obtained from spatial integration, and each provides useful information concerning the dynamics in the scene.

4.3.1 Streak Flow

Research in social behavior of pedestrians in crowds reveals that people tend to follow a pathway trailing pedestrians who have similar paths as a group [30]. As a pedestrian passes a point, there is a social expectation that any other pedestrian behind him/her would follow a similar path. Considering this social behavior, the actual, but invisible, flow of pedestrians has no gaps between individuals who are walking similarly. Hence, for crowd motion, gaps in the optical flow should be filled along trajectories with similar motion vectors prior to analysis.

In order to achieve an accurate representation of flow from crowd motion, we use the streaklines to compute a new motion field which we refer to as *streak flow*, denoted $\Omega_s = (u_s, v_s)^T$. To compute streak flow, we compute the streaklines by temporally integrating optical flow, as illustrated in Table 4.2, and forming the particles as in Equation (4.2). We describe the computation of u_s; computation of v_s is similar. Given data in the vector

$U = [u_i]$, where $u_i \in P_i, \;\; \forall i, p$, we compute the streak flow in the x direction at each pixel.

Based on equations (4.1), particle positions have sub-pixel accuracy. We compute a triangulation of pixels, which implies that each particle P_i has three neighboring pixels (nearest neighbors). At the sub-pixel level, it is reasonable to consider u_i to be the linear interpolation of the three neighboring pixels. Hence, we define

$$u_i = a_1 u_s(k_1) + a_2 u_s(k_2) + a_3 u_s(k_3) \;, \tag{4.3}$$

where k_j is the index of a neighboring pixel, and a_j is the known basis function of the triangulation of the domain for the j-th neighboring pixel. Using a triangular interpolation formula, each $u_s(k_i)$

is computed based on the relative positions of the three pixels and the particle. Using (4.3) for all the data points in U, we form a linear system of equations

$$Au_s = U \ , \tag{4.4}$$

where a_i are entries of the matrix A, and u_s is the least square solution of (4.4). [1]

Streak flows encapsulate motion information of the flow for a period of time. This resembles the notion of *particle flow* (equivalent to average optical flow) where advection of a grid of particles over a window of time provides information for segmenting the crowd motion. We argue that streak flows exhibit changes in the flow faster than particle flow, and therefore, they capture crowd motions better in a dynamically changing flow. This can be observed in Figure 4.3, illustrating sample frames from a video of a traffic intersection, which includes motions from both pedestrians and vehicles. The flow in the scene is unsteady and the different motion patterns appear in the video as the traffic lights change. The figure compares the streak flow to the particle flow and the optical flow in capturing temporal and local changes. For temporal changes the flow is compared at two different times: (1) At the start of the top-down flow of traffic (1st row), and (2) at the ending stage of the up-down traffic flow (2nd row).

Temporal changes: The first row of Figure 4.3 shows a frame from the sequence a few seconds after the change of a traffic light, so vehicles and pedestrians are now moving in a different direction, from top to bottom. By comparing the area to notice inside the red circle, it is evident that the streak flow is able to capture this change after only a couple of frames, but the particle flow

[1] www.mathworks.com/matlabcentral/fileexchange/8998-surface-fitting-using-gridfit

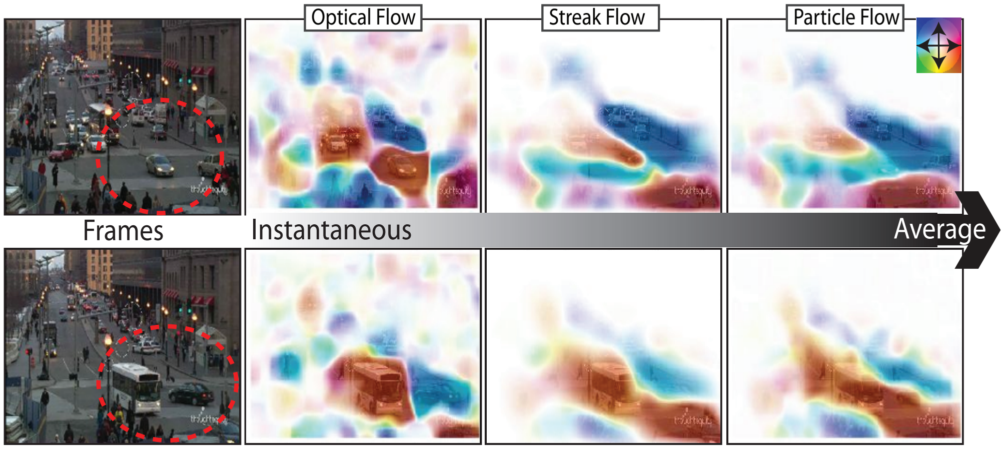


Figure 4.3: The comparison of optical flow, particle flow and streak flow for Boston sequence (color coded). The red circle indicates the area to notice.

lags in shaping to the new flow, and the optical flow shows choppy flow segments that are difficult to use for further analysis.

Local changes: Both streak flow and particle flow have the ability to fill in the gaps of the non-dense traffic flow. In second row of Figure 4.3, the optical flow shows the motion of a car making a left turn. The particle flow is unable to capture this change, and the region on the bus and car both show inconsistency compared to instantaneous flow. The figure shows that the streak flow was more accurate in exhibiting immediate flow changes over the car as well as the bus.

4.3.2 Potential Functions

Building on the fluid dynamics approach to crowd motion, we employ another concept from fluids providing a different point of view. In simplified mathematical models of fluids, it is often assumed that the fluid is incompressible, and irrotational. These assumptions imply several conservation properties of the fluid, but most importantly, they lead to potential functions, which are scalar functions that characterize the flow in a unique way. For this discourse, potential functions enable accurate classification of behaviors in a scene, which is not possible with streak flow alone. Application of potential functions to abnormal behavior detection is presented in Sections 4.4 and 4.5.

Since the optical flow $\Omega = (u, v)^T$ denotes a planar vector field, the Helmholtz decomposition theorem states that $\Omega = \Omega_c + \Omega_r$, where Ω_c and Ω_r respectively denote the incompressible and irrotational parts of the vector field. To clarify, an incompressible vector field is divergence free $\nabla \cdot \Omega = 0$, and an irrotational vector field is curl free $\nabla \times \Omega = 0$. Thus, there are functions ψ and ϕ, known respectively as the stream function and the velocity potential, satisfying $\Omega_c^{\perp} = \nabla\psi$ and $\Omega_r = \nabla\phi$ (see, for example [39]). Following [16], we use Fourier transforms to decompose incompressible and irrotational parts of the vector field and estimate the potential functions using

$$\begin{aligned}
\phi(x,y) &= \phi_0 + \frac{1}{2}\int_0^x (u_r(s,y) + u_r(s,0))\,ds + \frac{1}{2}\int_0^y (v_r(x,s) + v_r(0,s))\,ds\ , \qquad (4.5)\\
\psi(x,y) &= \psi_0 + \frac{1}{2}\int_0^y (u_c(x,s) + u_c(0,s))\,ds - \frac{1}{2}\int_0^x (v_c(s,y) + v_c(s,0))\,ds\ . \qquad (4.6)
\end{aligned}$$

Potential functions are computed in Corpetti et al. [16] and used in a meteorological application to track weather patterns in satellite images. In order to compute valid potential fields, one needs a dense motion field. In that particular application the motion fields are as dense as possible, but in crowd videos the degree of motion density can vary by large amounts. In addition, a potential function computed directly from optical flow is noisy with many valleys and peaks, which quickly

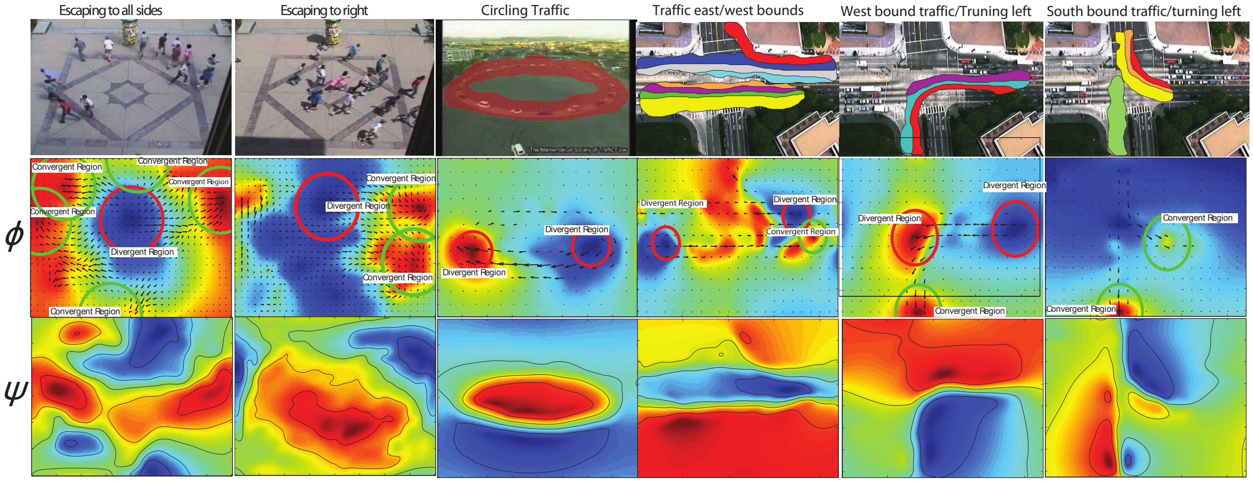


Figure 4.4: An illustration of discrimination power of potentials for six manually labelled behaviors. The first two columns, escape panic from UMN Dataset [2], column 3 shows circulating motion of cars in a lane, and columns 4 to 6 show traffic forming lanes from NGSIM dataset. Potentials are scaled to maximum value and plotted using jet colormap. (1st row) The lanes are overlaid with the frame for the steady motions. (2nd row) divergent regions (red circles) and convergent regions (green circle). (3rd row) Stream function and the related iso-contours are illustrated.

disappear and reappear. Streak flows enable us to compute reliable potential functions for crowd flow, incorporating local and temporal changes. In other words, we incorporate streaklines to compute smoothly evolving potential functions, which better reveal the dynamics of the crowd. In a broad view, the stream function ψ provides the information regarding the steady and non-divergent part of the flow, whereas the velocity potential ϕ contains information regarding the local changes in the non-curling motions. Moreover, to have a complete picture of the flow we need information from both potential functions. With this perspective, we illustrate the strength of potentials in discriminating lanes and divergent/convergent regions in five different scenes in Figure 4.4. In this figure, the velocity potential is accountable for capturing unsteady changes in the flow. For instance, escape to the sides of the scene corresponds to a valley in the center of ϕ and formation of surrounding peaks on the sides. Furthermore, the stream function ψ is incorporated to detect lanes in the steady motion of vehicles. The area between contours of ψ (i.e., streamlines) show the regions of steady and non-divergent motion such as lanes. The algorithm for detection of lane and divergent/convergent regions is explained in Section 4.4.

4.4 Applications of Streaklines

Using streak flow and potential functions, we demonstrate the strength of our approach for crowd segmentation and abnormal behavior detection in unsteady flows. In the end, we find that our method performs better than other methods for solving these problems.

4.4.1 Crowd Segmentation

In this algorithm, we segment every frame of the video into regions of different motions based on the similarity of the neighboring streaklines. Similar streaklines correspond to similar trajectories of particles passing from neighboring pixels over a period of time. Hence, it captures the affinity of current and previous motions at these pixels. Figure 4.5 presents the block diagram of the segmentation algorithm. First, frame by frame optical flow of the video is computed. Using the optical flow, a set of particles are then moved over the frame to construct the streaklines and the streak flow. These quantities are used to compute similarity in a 8-connectivity neighborhood. For every pair of pixels i and j, the similarity is computed in terms of streaklines and streak flow.

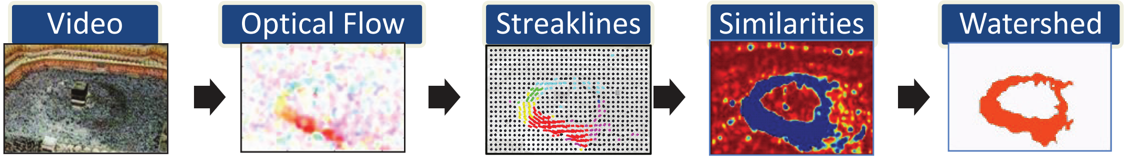


Figure 4.5: The crowd segmentation algorithm.

Each pixel is associated with a streakline of length l. The streakline similarity is computed using the sum of the normalized projections of internal vectors as $R_s(i,j) = \sum_{m=0}^{l-1} prj(X_m^i, X_m^j)$, where X_m^i and $prj(\cdot,\cdot)$ are defined in Figure 4.6.a. Streak flow similarity is computed as $R_\Omega(i,j) = |\cos(\angle\Omega_s^i) - \cos(\angle\Omega_s^j)|$, where $\angle\Omega_s^i$ is the angle of the streak flow vector at pixel i. In order to define boundaries of the regions, we compute the similarity map at every pixel using

$$H(i) = \sum_{j \in N(i)} \alpha R_s(i,j) + \beta R_\Omega(i,j) \ , \tag{4.7}$$

where α and β are weights regulating the share of streakline and streak flow similarities in the final segmentation. We use $\alpha = 0.8$ and $\beta = 0.2$ in the experiments. Since similar motions over time build similar streaklines and streak flows, boundaries of different motions form valleys in the similarity map. Using the negative of the similarity map, we segment the crowd into regions of similar motion with watershed segmentation. Results are presented in Section 4.5.1.

Lane detection: In addition to segmenting a frame into regions of consistent motion, we combine information from potentials to detect lanes in each segment. As stated in section 4.3.2, the area between contours of ψ corresponds to the steady flow, and the rate of the incompressible flow between a pair of contours is equal to the difference between the values of ψ on those contours.

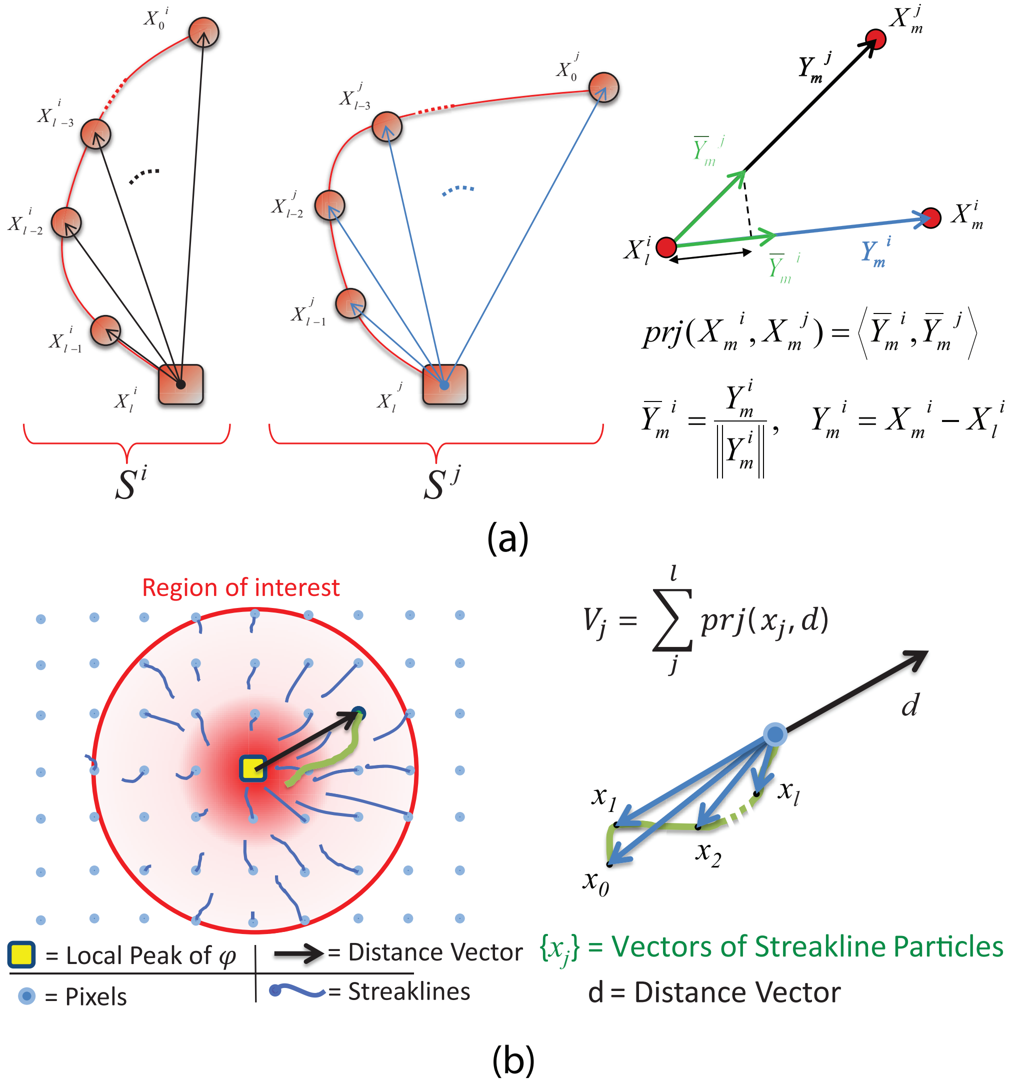


Figure 4.6: (a) Streaklines S^i and S^j are sets of vectors $X^i_{1..L}$ and $X^j_{1..L}$. The originating point of streaklines (rectangle), the particles (circles), and the normalized projections of the vectors are used for computing the similarity of streaklines. (b) The computation of divergence factor, V_i, for a region of interest.

Considering this, we detect lanes as parts of a segmented region that fall between two contours of the stream function by a simple intersection operation (see Figure 4.4).

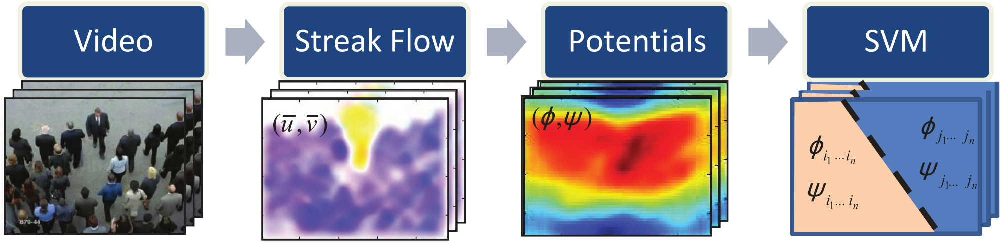


Figure 4.7: The abnormal behavior detection algorithm.

4.4.2 Abnormal Behavior Detection

To detect abnormal behavior of crowds, it is necessary to have a global picture of the behavior in a scene, for which we use potential fields. The surfaces ϕ and ψ characterize particle positions and velocities in a global sense, and abnormal behaviors are simply detected as large deviations from the expected. Here, we present an algorithm to detect abnormal behavior in crowds using potential functions for the flow.

Figure 4.7 shows the block diagram for the algorithm. For every frame in a video sequence, the Streak flow $\Omega_s = (u_s, v_s)^T$ is computed, and the potential functions of the frame $\{\phi, \psi\}$ are computed using equations (4.5) and (4.6). The peaks and valleys of the potential surface convey information regarding the global behavior of the flow (Figure 4.4). Thus, potentials provide new features to distinguish global behaviors in the crowd in compact form. For every frame, a feature vector V is formed by concatenating the values of ϕ and ψ of that frame. Using feature vector V, we recognize behaviors in each frame by training a support vector machine (SVM) classifier. In Section 4.5, we provide comparative results of abnormal behavior detection using potentials.

In addition to detecting abnormal behaviors, we incorporate streaklines and the velocity potential ϕ to provide a description of the anomaly based on divergent/convergent regions. The extrema on velocity potentials correspond to divergent or convergent regions. To robustly detect these regions, we find the major local extrema of ϕ, and then compute the average *divergence factor*, $\bar{V} = \frac{1}{n}\sum_i V_i$, where V_i is defined in Figure 4.6.b, and n is the number of pixels in the radius r of the extremum point. Simple thresholding of this factor distinguish divergent/convergent regions as

$$Region\ Type = \begin{cases} Divergent, & \text{if } \bar{V} > T \\ Convergent, & \text{if } \bar{V} < T \end{cases}. \tag{4.8}$$

In the experiments, r is set fixed empirically for each scene and $T = 0$. As it is illustrated in Figure 4.4 the escape panic scene involves the divergent region in the center and convergent regions on the sides to which the crowd is running. Similarly, a sudden change in the direction of turning vehicles or the entry/exit points form divergent/convergent regions. The circular regions in the second row are the actual output of our algorithm. Obviously, there are some mistakes (around 20%). For example, in circling traffic, column 3, the region on the left is detected incorrectly as a divergent region whereas the majority of the vehicles are entering to that region.

4.5 Experimental Results

We present results of algorithms outlined in Section 4.4, using experiments on two datasets. A stock footage dataset from the web [47] is used for streakline analysis, and a dataset from the University of Minnesota [2], which contains 11 videos of crowd escape panic, is used to evaluate the effectiveness of potentials for abnormal behavior detection.

4.5.1 Results of Crowd Motion Segmentation

Results of our proposed segmentation algorithm are provided here. We compare with the state of the art [5], considering crowds with dynamic segmentations, such that the motion patterns vary in time exhibiting different states of behavior.

Figure 4.8 provides segmentation results for two scenes, and video frames are overlaid by colored segmentation regions. In this experiment, the length of streaklines and pathlines is $l = 40$. On the left side of Figure 4.8, an intersection is shown in Boston, containing three behavioral phases represented by frames 40, 197, and 850. (1) South bound traffic is formed. (2) Traffic lights change and an east/west bound (from/to station) a flow of pedestrians emerges. (3) Traffic lights change again, and a north bound vehicle flow is formed together with an east bound pedestrian flow. On the right side of Figure 4.8, an intersection is shown in Argentina containing three behavioral phases. (1) East/west bound traffic is formed. (2) After the traffic lights change, a south bound vehicle flow and a north/south pedestrian flow develop. (3) Traffic lights change to the first phase and east/west bound flows resume. Frames 115 and 213 illustrate the start of phases 2 and 3, respectively. The optical flow of this video is particulary noisy as it is based on time-lapse imagery, whereas the Boston sequence is a regular $30fps$ video. Videos are available in the supplementary material.

Figure 4.8 demonstrates the segmentations based on streaklines are spatially and temporally pronounced and more accurate in dynamic scenes than the state of the art. We highlight the gains in using our method in each frame: (Frame 40) A walking pedestrian and the north bound vehicle

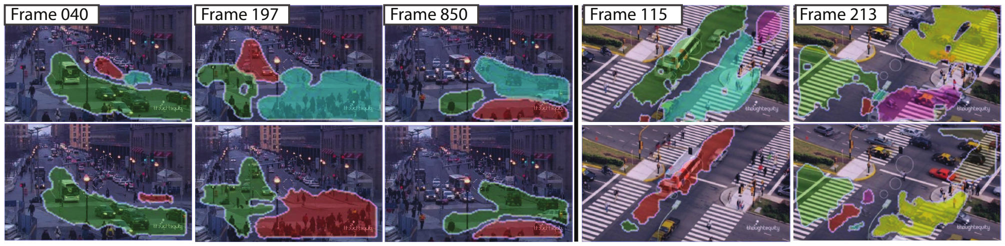


Figure 4.8: The comparison of segmentation results using streaklines (1st row), and pathlines [5] (2nd row) for scenes with unsteady motions.

motion are segmented correctly. (Frame 197) Pedestrians are distinguished from the south bound cars. (Frame 850) A south bound pedestrian (first row, green) is separated from north bound vehicles. (Frame 115, 4th column) Different pedestrian flows are distinguished (first row, cyan and purple). (Frame 213) West bound vehicle flow (first row, yellow) is segmented earlier, at the start of phase 2 of the video.

In Figure 4.9, the quantitative comparison of the proposed segmentations method and [5] is provided. In this experiment, frame by frame segmentations of both methods are compared as follows. The number of objects (human/vehicle) in each segmented region is counted provided that its direction of motion is no more that 90 degrees apart from the direction motion of the majority of the objects. We refer to this number as the number of correctly segmented objects (see Figure 4.9.a). To evaluate the methods, this number is counted manually for a subset of frames of Boston and Argentina video sequences. Figure 4.9 demonstrates that streakline segmentation outperforms the state of the art in number of correctly and incorrectly segmented objects.

4.5.2 Results of Abnormal Behavior Detection

This section illustrates results for abnormal behavior detection on the UMN dataset [2], containing 11 sequences for 3 scenes. In this dataset, pedestrians initially walk randomly, and exhibit escape panic by running in different directions in the end. Figure 4.4 shows that potential functions provide rich information about global behavior. Interesting properties of potentials are revealed as we compare ϕ for frames where people escape to all sides to the frames for which people run in a single direction (2nd column).

In order to illustrate the strength of potentials in representing the global behavior we compared our method using different features. In *experiment (a)*, we first use frame-based potentials as the input features V for training a SVM with RBF kernels. Second, we use vectorized streak flow $\Omega_s = (u_s, v_s)$ and third, we use average baseline optical flow (pyramidal LK) to perform the same task. Figure 4.9.e compares the recognition results using any of these three features for a different number of training examples. In order to reduce the computation time, we downsample the features of each frame by factor of $n = 20$. In this experiment, the frames from different scenes in the dataset are combined in a single pool and a portion is selected as the train set and the rest is considered as the test set (no overlaps). The figure shows that after increasing the number of examples to merely 20%, the potentials show considerable improvement in performance. In addition, the figure illustrates the strength of streak flows compared to particle flow in providing information for abnormal behavior detection.

In *experiment (b)*, we performed a leave-one-out cross validation on the UMN dataset using downsampled versions of potentials and average optical flow. In this experiment, we trained a SVM with RBF kernels on 10 videos and computed the false positive and true positives on one video sequence and repeated this for all the 11 videos. Figure 4.9.f illustrates the ROC of this experiment which indicates improvement using potentials over baseline optical flow.

4.6 Conclusion

Based on a Lagrangian particle dynamics framework for fluid flow, we juxtapose three vector field representations of the flow, given by streamlines, pathlines, and streaklines. With application to problems in segmentation and abnormal behavior detection for crowd and traffic dynamics, we show that the streakline representation is advantageous. When compared to the other two representations, which are commonly used to solve problems in computer vision, streaklines demonstrated the ability to quickly recognize temporal changes in a sequence, in addition to finding a balance between recognition of local spatial changes and filling spatial gaps in the flow. When used to compute potential functions and to perform segmentation, the streakline approach was superior to using optical flow and comparable to using particle flow, aside from the ability to recognize scene changes. With regard to abnormal behavior detection, the method of streaklines proved superior to both of the other representations, and the introduction of potential functions for this purpose proved valuable.

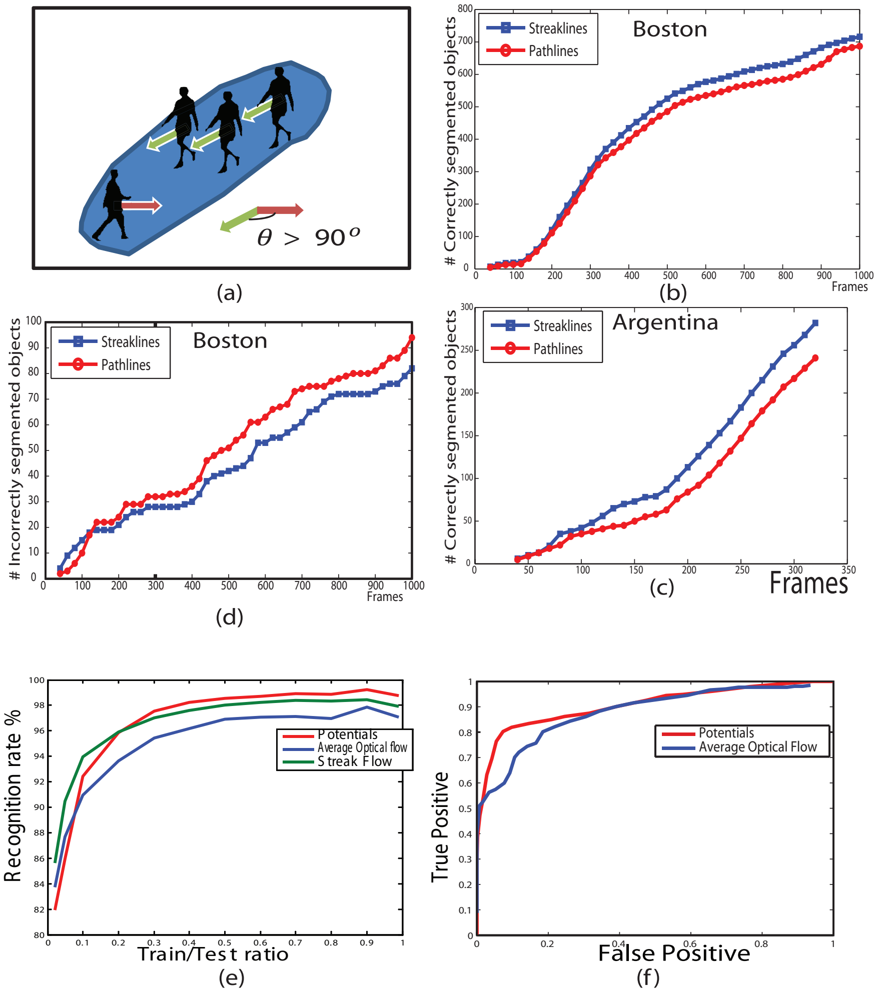


Figure 4.9: (a) The criterion for segmentation evaluation, (green) correctly segmented object, (red) incorrectly segmented object. (b), (c), and (d) Quantitative comparison of segmentation results using streaklines (blue), and pathlines [5] (red). (e,f) Abnormal behavior recognition, (e) Variation of the number of training examples. (f) ROC of the cross validation.

CHAPTER 5: GROUP BEHAVIOR RECOGNITION

5.1 Introduction

In the previous chapters, we investigate the crowd behavior as a single holistic entity which undergoes certain dynamics or as a body of parts where each exhibits a coherent but independent motion. In this chapter, we propose a new approach to model the inter-relations of motion entities in a scene. Therefore, a group behavior is modeled not as isolated motions patterns but as coordinated movements of people. The intuition comes from the fact that behaviors are a mixture of coordinated but distinguishable motion patterns in certain geometrical and temporal order. Thus, without studying the relation of the motions patterns in a scene, any behavior recognition algorithm will face severe scalability issues when applied to another scene or in a different time other than the place and time it is designed for. In this chapter, we propose a method to represent the spatial layout of the motion patterns in a compact and discriminative form. We propose a method to model the structural layout of the behaviors by considering the higher order inter-relation of motion descriptors in the scene. We apply this method to solve the new problem of group choreography recognition in conjunction with introducing a new dataset.

5.2 Group Behavior Recognition Challenges

The group behavior is a challenging task in computer vision. The ideal method to perform this task should address the following concerns.

- Group behaviors, such as group dances or stage performances, involve a large variety of articulated movement in addition to the common challenges such as occlusion and clutter. Therefore, the motion representation should be robust and discriminative.
- The group behaviors are collections of different actions in certain geometric layout or temporal order. This creates a large variance in the space of possible behaviors even with a limited number of possible actions. Therefore, the behavior recognition method should be scalable and comprehensive.

To study the group behaviors of humans, we focus on choreography recognition since the popularity of the subject has created an abundance of data on Internet which definitely helps us in creating a dataset and in applying machine learning algorithms. However the concept and the structure of the problem is common among other group activities, and researchers can apply the same proposed techniques in other contexts such as transportation research, group sports like soccer and basketball, stage performances, religious ceremonies, and even animal swarms with proper adaptations.

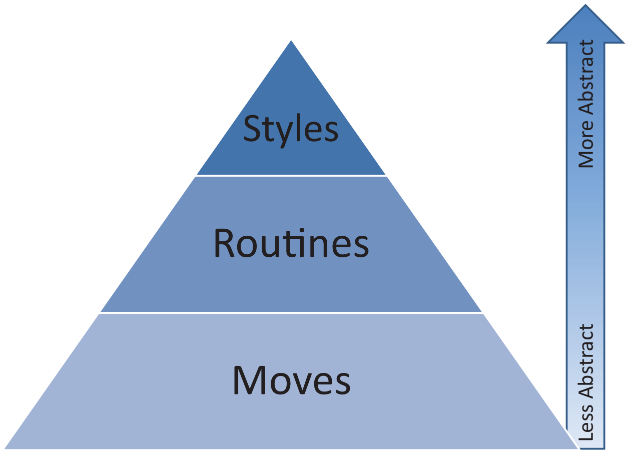


Figure 5.1: The hierarchy of the choreography concepts.

5.2.1 Elements of Choreography

Choreography is the art of dance design and the word choreography literally means *dance writing* in Greek language. The role of choreography in dance performances is the same as the role of musical notation in creation of music. In an inclusive definition, a choreography is a composition of pieces of visual manifestations such as movements, light, space, music, costumes, and improvisations which are referred by *dance elements*. In this thesis we are investigating the choreography formation merely based on composition of movements or the motion. Therefore, in the following arguments we plainly consider the movement as the only element of the dance. The further exploration of visual understanding of chorographies based on other elements or their combination is left for future research and is beyond the scope of this thesis.

Dominantly, the building blocks of the dance movements are called *dance moves* or *dance steps* which are short or atomic, yet articulated, actions in tune with music. There are multitudes of dance moves such as Michael Jackson's Moonwalk, basic step, and pomp turn [15]. Table 5.1 contains examples of dance moves, and Figure 5.2 illustrates sample dance moves. The hierarchy of concepts in choreography is depicted in Figure 5.1. The dance moves are at the lowest level of semantic hierarchy of choreographies.

The mid level concept is called a *dance routine*, which is a series of *dance moves* in certain composition. For instance, a choreographer would design a dance routine by composing three dance moves such as: basic step, turn, and basic step. The composition, definitely, encompasses the duration and the rhythm of the movements as well, but we ignore that for the sake of argument. Based on our observations, the dance routines are much easier to visually recognize than atomic moves.

The highest conceptual level of understanding dances is the *dance style*. The style of a dance is defined by the composition of the dance routine. For example, in a certain style of dance the choreographer would use only certain moves in certain spatio-temporal formations which is the signature of that style. Table 5.1 enlists a few popular dance styles such as Ballet, Hip hop, and Country dance. Certain styles of dances are performed in groups and the choreography for those dances contains the layout and the inter-relation of dance routines of individuals or subgroups in the whole performance. In this thesis we focus on group dances in the following styles: (1) Bollywood (Hindi), (2) Country, (3) Folk, and (4) Line.

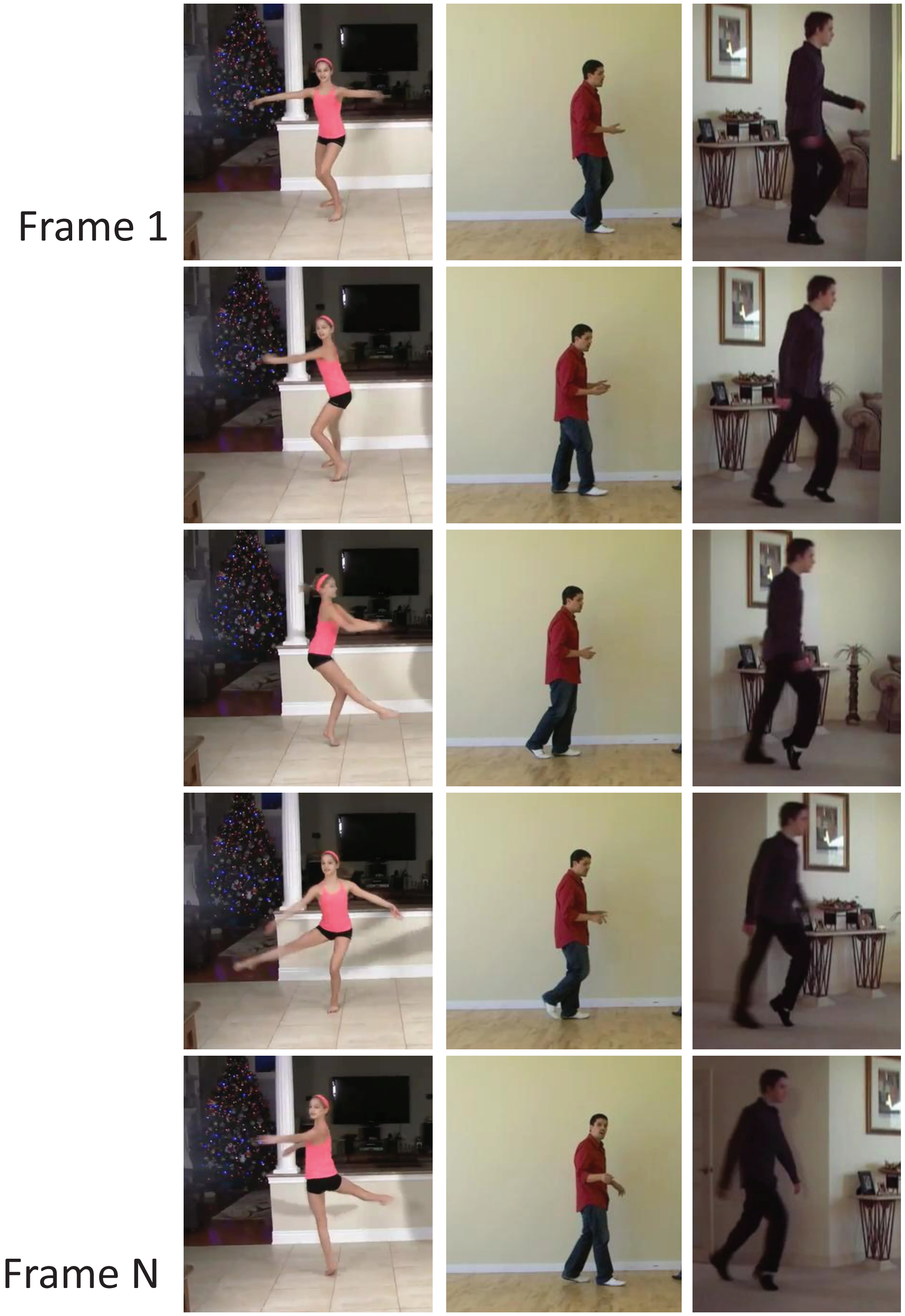


Figure 5.2: Sample dance moves. (Left) ballet pomp jump, (middle) Basic steps of Salsa, and (right) Moonwalk.

Dance Style	Common routines	Common Moves	Group Dance
Ballet	Adagio Allegro Coda	Pas de chat Pas de cheval Pas de valse	Yes
Hip hop	Breaking Locking Popping	Running man Kick cross step The wave	Yes
Country	Chass Clockwise Gypsy Stars	Walk Turn Bending arms Holding hands	Yes

Table 5.1: The list of a few popular dance styles with their common routines and moves.

5.3 Coordinated Behaviors as Bag of Phrases

In this chapter, we introduce an algorithm to model the group behavior involving coordinated behaviors in the bag of words framework. However, instead of only using motion descriptors such as Dollar motion descriptor [19] to construct the code book, we consider the first, second, third, and other higher order combination of visual words. Even tough the framework presented in this chapter is focused toward group behavior recognition, the concept is general and can be applied to other computer vision problems that involve bag of visual words representation and the spatial configuration of features is important. Visual words are quantized motion descriptors from the set of all visual words or the visual word vocabulary (codebook). A group of visual words in certain geometrical layout is referred to as a *visual phrase*. Our proposed method extends the recent work of [77] in the object recognition research community in using higher order visual features that contain the spatial layouts of visual words, and applies this extension to video activity recognition.

Despite the success of the Bag of Visual Words (BoVW) model in recognizing objects, actions, and activities, it has a notorious limitation in ignoring the spatial relationships among the visual words. This plays as a major disadvantage in modeling objects or behaviors with distinct spatial structures, and therefore, reduces the discriminative power of the method. For instance, in an activity recognition problem such as group dancing, bag of visual words cannot discriminate between distinctly different behaviors which occur in different spatial layouts but involving the same visual words. Figure 5.3 illustrates this ambiguity in visual word model is an example of group behavior where two different behaviors (a and b) would, incorrectly, have similar representations. Moreover two similar behaviors (c and d) would, ambiguously, have totally different representations.

In this section, we introduce an extension to the recent work in [75] to identify visual phrases invariant to any geometric transformation such as similarity, affine, and projective transformations. In addition, we introduce an efficient algorithm for the special case of similarity transformations. Finally, we apply the extended representation to the choreography recognition problem. We follow the notation of the original framework for geometrically preserved bag of visual phrases in [75], and we present our adaptation of the method to activity recognition in videos as our first step toward extending it.

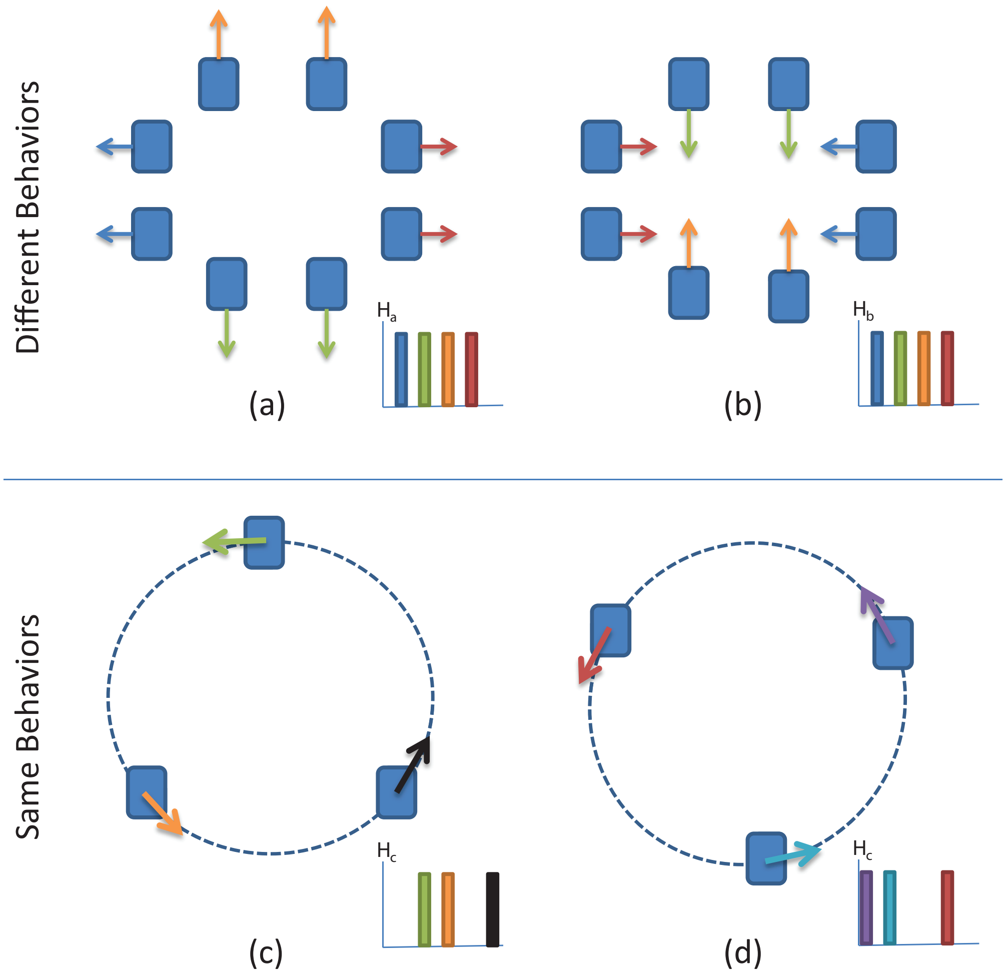


Figure 5.3: A schematic of the ambiguity of group behaviors in bag of visual words representation. The spatio-temporal volumes are represented by rectangular patches and the arrows indicate the visual words. The color arrows represent the label of the visual word in the codebook. (a) and (b) illustrate two different behaviors, convergence and divergence, which is confused in BoVW as the same behavior. (c) and (d) illustrate two similar behaviors, circling, which have totally different BoVW representations.

5.3.1 Adapting Visual Phrases for Videos

In this section, we review a general definition of visual word and visual phrases methods by adapting the core idea to videos and activity recognition problems. A visual phrase of length k is defined as k visual words in a certain spatial layout. Different words or different spatial layouts define different phrases. We first review the simple case of visual phrases of length one which the representation reduces to visual words. In bag of visual words framework, an image or a video clip is represented by a histogram of visual entities or *visual words*. The visual words are quantized sets of spatial pathes or spatio-temporal volumes of data using a clustering algorithm such as k-means algorithm. Given an initial visual vocabulary $V = \{w_1, w_2, ..., w_m\}$, where $w_i \in R^d$, and d is the dimension of the descriptor for the patch or the spatio-temporal volume, a video clip c is represented by a histogram $\Phi = \{h_1, h_2, h_3, ..., h_k\}$ where h_i is the term frequency of visual word h_i in the video clip. Later in this section, we use a more comprehensive representation of a video clip in the bag of visual phrases representation.

Kernel methods, such as Support Vector Machines (SVM), are among the most popular approaches for developing learning algorithms for activity recognition using BoVW representation. A kernel function is a function that calculates a metric, usually an inner product, between two patterns after mapping to the feature space. The mapping, in this case, is the visual word representation of a video clip or $\phi(c)$. For any mapping $\Phi : C \to F$, from the input space C, to the feature space F, a kernel function is defined as

$$K(c_i, c_j) = \langle \Phi(c_i), \Phi(c_j) \rangle, \forall c_i, c_j \in C. \tag{5.1}$$

However, in BoVW framework the kernels does not capture any spatial information, and the popular kernels such as histogram intersection and $\chi - square$ ignore the geometrical relationships of the visual words.

In order to overcome these shortcomings, in [75, 76], the authors present a new kernel for the bag of visual words which considers visual phrases in addition to visual words. The basic idea is to compute the metric not only based on the visual words but also based on their spatial locations and geometrical relationships. The trivial approach would be to redesign the mapping function Φ to include the histograms of visual phrases as well. Then, the previously known kernels can be applied directly. However, this approach is extensively costly in terms of memory and speed as it increases the complexity of the algorithm exponentially.

Therefore, in [75, 76] an efficient method is proposed to construct the kernel based on the matching co-occurring visual phrases using an intermediate transformation which is Generalized Hough Transform (GHT). Figure 5.4 illustrates the basic idea. We refer to the visual words or visual phrases as *features*, and in the following, we describe the video representation and the kernel for higher order spatial relations of features in the framework presented in [76].

Video representation: A video clip c is represented as set of pairs of visual words x_i and its corresponding spatial information in the frame. That is, $c = \{(w_1, r_1), (w_2, r_2), (w_3, r_3), \ldots, (w_n, r_n)\}$, where $r_i = (x, y, s)$ is a triplet of xy location in pixels and the scale of the feature.

Kernel for higher order Spatial Relations: A feature of order n is defined as a combination of n visual words in a certain spatial relationship as is referred to by f^n. Therefore, different relative spatial locations or scales of the features creates different features. For the visual vocabulary

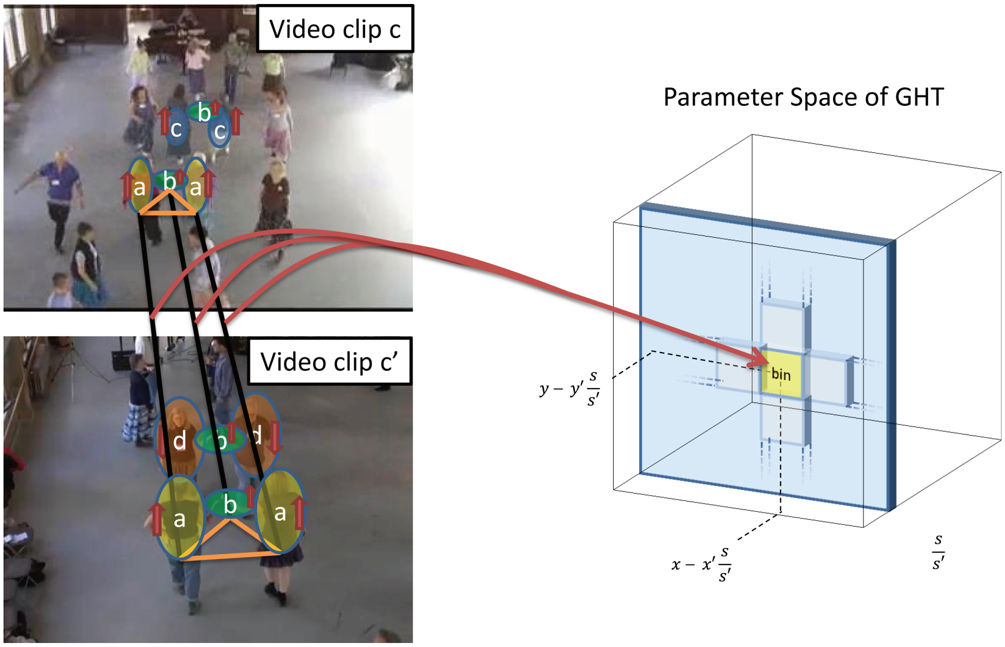


Figure 5.4: Illustration of the algorithm to count the co-occurring visual phrases which are in certain geometrical layout in two video clips using Generalized Hough Transform. (a) Two frames of two different video clips are illustrated. Ellipses illustrate the visual words and the color and the labels refer to the word index. The size and orientation of visual words are represented by the size of the ellipse and the orientation of the major axis. The yellow triangle indicates the spatial relationship between three words. (b) The parameter space of GHT that considers translation and scale between visual words. Each word correspondence contributes to a bin in the parameter space. A group of words in the same spatial relationship contributes to the same bin.

$V = \{x_1, x_2, ..., x_m\}$, features of order n are members of a set of $S_n = m^n \times q$ selections of n visual words in q possible spatial layouts. n-word feature representation of a video is presented as $\Phi_n = \{h_1, h_2, h_3, \ldots, h_{S_n}\}$, where h_i is the number of occurrence of the corresponding feature in the video clip c. Therefore, for any mapping $\Phi_n : C \rightarrow F^n$, from the input space C, to the $n - th$

order feature space F^n, a kernel function is defined as

$$K_n(c_i, c_j) = \langle \Phi_n(c_i), \Phi_n(c_j) \rangle, \forall c_i, c_j \in C. \tag{5.2}$$

By combining the kernels from different orders of features, the final kernel is written as

$$K(c_i, c_j) = \sum_{r=1}^{n} \alpha_r K_r(c_i, c_j), \tag{5.3}$$

where $\alpha_r = \mu^{1-r}$ and $0 < \mu < 1$ weighs the higher order features.

Defining the kernel as an inner product in the feature space facilitates its use for SVM learning as it satisfies the Mercer's condition. However, direct computation of the kernel is computationally expensive because the size of feature space grows exponentially with respect to n, the order of the features. It is shown in [75] that we can avoid facing the computation of $\Phi(c)$ by computing the inner product indirectly and efficiently. This is possible because the inner product equals to the sum of the co-occurrence of all $n - th$ order features. Therefore, the practical solution is to develop a method to count the number of features co-occurring in both video clips c_i and c_j in the similar spatial layouts. In the next section, we review the use of Generalized Hough Transform for this purpose.

5.3.2 Generalized Hough Transform and the Co-occurring Features

As described in [75] and [76], the Generalized hough Transform (GHT) is effective and efficient in computing the number of co-occurring n-th order visual features. As illustrated in Figure 5.4, the algorithm forms a 3-dimensional parameter space T of possible transformations using (x, y) coordinates and scale s. The space is quantized into a finite number of bins to allow invariance to small deformations. Every bin in this 3-dimensional space corresponds to a specific amount of possible translation and scale. A pair of spatio-temporal features with the same visual word assignments ($w_i = w'_i$) from two video clips c and c' contributes a single vote to a bin with corresponding coordinates $(\lfloor x_i - \frac{s_i}{s_i} x_i \rfloor, \lfloor y_i - \frac{s_i}{s'_i} y'_i \rfloor, \lfloor log(\frac{s_i}{s'_i}) \rfloor)$. The value of the corresponding bin is accumulated by considering the location and scales of every matching visual word pairs from two video clips. After casting all the votes from all the possible pairs, the value at each bin, R, indicates the total number of pairs with a specific relative transformation. Therefore, the number of co-occurring features of order n is represented by the number of n chooses from the R possible choices. This equals to $\binom{R}{n}$ for any $n <= R$.

Therefore, to compute the $n-th$ order kernel K_n, the kernel is computed as

$$K_n(c_i, c_j) = \sum_{b=1}^{Q} \binom{R_b}{n}, \tag{5.4}$$

where Q is the total number of bins in the parameter space and R_b is the number of votes in b-th bin.

5.3.3 Difficulties in Matching Visual Phrases Invariant under Similarity Transformations

Extending the method in previous section to include invariance under rotation, scale, and translation (similarity transform) is not a trivial task. First, adding another parameter, rotation, to the parameters space grows the space exponentially and significantly affects the speed of the algorithm and raises serious concerns regarding required memory space which makes the method impractical. In [76, 77] authors ignore the rotation because of the same concern. Second, the parameters space is the space of similarity transform which belongs to a nonlinear manifold. Quantizing this space into fixed-size bins leads to inconsistent matchings in regards to similar transformations, since the fixed-size bins ignore the metric of the manifold. To avoid these two problems the authors in [75] proposed a peudo-invariant method that includes rotation and scale, which we will discuss its limitations in the following.

A close scrutiny reveals that the method proposed in [75] incorrectly matches visual phrases in degenerate cases. That is because, in their method, the authors normalize the scale or the rotation of features separately before comparing the spatial distances, and they do not include scale in the parameter space. However, this creates incorrect matches of high order visual phrases. In addition, the number of possible degenerate cases increases exponentially with increase in the feature order. Therefore, the higher order features will be affected even more. Figure 5.5 illustrates an example degenerate case where a triplet of features is matched using the method in [75] even though the geometric layout is totally different. In this figure, three features have occurred, where the scale of one of them, feature b (in green color), has changed in two different video clips. Therefore, the geometric layout of the triple is different in these video clips. In other words, no single geometrical transformation, involving translation and scale, exists that could explain the transformation of pair of features from the first clip to the second clip. Therefore, the three features in these two video clips do not belong to a same visual phrase. This suggests that to perform the higher order feature matching which is invariant to geometric transformations, we need to consider the components of a geometric transformation altogether. In the next section, we propose a method that achieves this goal and correctly matches high order visual words under any geometrical transformation.

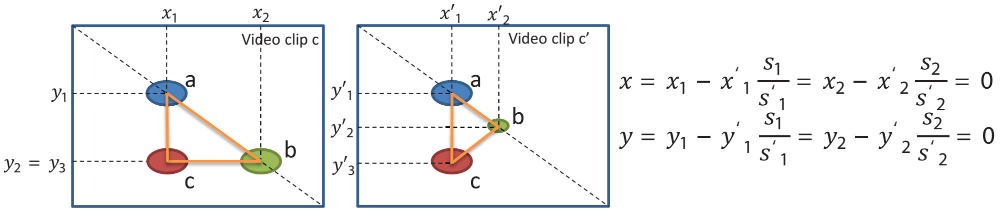

x_1
x_2
Video clip c
y_1
$y_2 = y_3$
a
b
c
x'_1
x'_2
Video clip c'
y'_1
y'_2
y'_3
a
b
c
$x = x_1 - x'_1 \frac{s_1}{s'_1} = x_2 - x'_2 \frac{s_2}{s'_2} = 0$
$y = y_1 - y'_1 \frac{s_1}{s'_1} = y_2 - y'_2 \frac{s_2}{s'_2} = 0$

Figure 5.5: An example of a degenerate case where the method in [76, 77] incorrectly matches a third order feature from two video clips. Visual words are identified with colors and labels. (a) The first instance of the visual features, $x_1 = \frac{1}{2}x_2 = x_3, y_1 = \frac{1}{2}y_2 = \frac{1}{2}y_3, s_1 = 2s_2 = s_3$. (b) The second instance of the visual features, $x'_1 = 2x'_2 = x'_3, y'_1 = y'_2 = \frac{1}{2}y'_3, s'_1 = 2s'_2 = s'_3$.

5.3.4 High Order Features under Similarity Transformations

In this section, we introduce a novel and efficient method to compute the kernel for high order features that is invariant under geometrical transformations. We focus on similarity transformations (i.e., translation, rotation, and scale) in this section, since they are using an efficient algorithm to compute the metric of similarity transformation in the nonlinear manifold. However, excluding the efficient metric computation, the proposed framework is general and applicable to any arbitrary matrix transformation such as affine or projective transforms. In our proposed method, since we consider the correct metric of the nonlinear manifold of the geometrical transformations, we are able to form a smaller parameter space with far less number of quantized bins and use soft assignments instead of hard assignments (voting) to form the parameter space for the Generalized Hough Transform efficiently. In this way, our method gains performance over the previous work by maintaining the same order of computation complexity.

The parameter space of the similarity transforms has four components: (x, y, s, θ). We quantize the parameter space such that each dimension has q levels. Therefore, the total number of bins in the parameter space is q^4. Each bin in the parameter space corresponds to a quadruplet of x-translation, y-translation, scale, and rotation parameters. In other words, a similarity transformation is a way to describe the quadruplet. We represent each bin, i, in the parameter space as $\mathbf{T}_j = \mathbf{R}(\theta)\mathbf{S}(s)\mathbf{T}_r(x, y)$, where $\mathbf{T}(\theta)$ is the rotation, $\mathbf{S}(s)$ is the scale, and $\mathbf{T}_r(x, y)$ is the translation transforms, and the ranges of the parameters are

$$x \in [0, 1], y \in [0, 1], s \in [s_{min}, s_{max}], \theta \in [-\pi, \pi], \tag{5.5}$$

where the (x, y) location is normalized with width and height of the frame. Provided the bins, we extend the video clip representation, c, to include the visual words w_i and its corresponding location, scale, and orientation in the frame. That is $c = \{(w_1, r_1), (w_2, r_2), (w_3, r_3), \ldots, (w_n, r_n)\}$, where $r_i = (x, y, s, \theta)$. Then, in our proposed method, we form a similarity transformation $\mathbf{T}$ in homogeneous coordinates that corresponds to every pair of spatio-temporal volumes of data with the same visual word assignments ($w_i = w_i'$) in two video clips c and c' such that

$$\mathbf{T} = \begin{pmatrix} s \times cos(\theta) & -s \times sin(\theta) & t_x \\ s \times sin(\theta) & s \times cos(\theta) & t_y \\ 0 & 0 & 1 \end{pmatrix}, \tag{5.6}$$

where $s = \frac{s_i}{s_i'}$, $\theta = \theta_i - \theta_i'$, and $(t_x, t_y)^T = (x_i, y_i)^T - (cos(\theta), sin(\theta))(sx_i', sy_i')^T$.

Given the similarity transformation matrix $\mathbf{T}$ for a pair of feature points, the bins in the parameter space are accumulated by weighted votes. These votes are computed with respect to the distance of $\mathbf{T}$ with the corresponding transformation to the center of the bin such that the bins closer to $\mathbf{T}$ get higher shares of the vote. Hence, the vote for the bin j is computed as

$$\beta_j = \frac{e^{-d(T,T_j)/\lambda}}{\sum_j e^{-d(T,T_j)/\lambda}}, \tag{5.7}$$

where $d(.,.)$ is the distance defined on transformation matrices. After accumulating the votes of all of the pairs we use integer part of values in the bins. Since, our extension has merely changed the vote casting method, the rest of the GTH method, which is the computation of the kernel for $n-th$ order features, remains the same as Equation (5.4). Figure 5.6 illustrates the concept.

We know that geometric transformation matrices belong to a nonlinear manifold [66], and the distance on that manifold is, therefore, a non-Euclidean metric. The distances on manifolds are defined in terms of minimum curves between points on the manifold [23]. The curve with the minimum length is referred as geodesic and the length of the curve is the intrinsic distance. The intrinsic distance between two similar transformations is, therefore, defined as

$$d(T_i, T_j) = \parallel log(T_i^{-1}T_j) \parallel_F, \tag{5.8}$$

where $log(T) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} (T-I)^i$ is the matrix logarithm and $\parallel . \parallel_F$ is Frobenius norm. Provided the metric of the geometrical transformations, we are able to use GTH method for the geometric layout of visual features.

The computation of $d(T_i, T_j)$ for arbitrary transformations can be a bottleneck as it involves matrix logarithm which is typically calculated through an iterative algorithm. However, for special cases of pure rotation or similarity transformation, direct solutions exist which boost the computation of matrix logarithm. We incorporate the method in [22] to compute the logarithm of the transformation matrix as

$$\log(\begin{pmatrix} s\mathbf{R} & \mathbf{v} \\ \mathbf{0} & 0 \end{pmatrix}) = \begin{pmatrix} \sigma\mathbf{I_3} + \mathbf{A} & \mathbf{P^{-}1v} \\ \mathbf{0} & 0 \end{pmatrix}, \tag{5.9}$$

where $\sigma = ln(s)$, $A = \log((T_r)(\theta))$, v is the translation vector, and s is the scale. Since the rotation matrix $\mathbf{R}$ is skew symmetric, its matrix logarithm can be computed through Rodrigues' formula [48] as

$$A = \log(\mathbf{T_r}(\theta)) = \begin{cases} \mathbf{0}, & \theta = 0 \\ \frac{\theta}{2\sin\theta}(\mathbf{T_r}(\theta) - \mathbf{T_r^T}(\theta)), & \mid \theta \mid \in (0, \pi) \end{cases}. \tag{5.10}$$

$\mathbf{P}$ is defined as

$$\mathbf{P} = \frac{e^{\sigma}-1}{\sigma}\mathbf{I_3} + \frac{1}{\theta^2+\sigma^2}(\frac{\sigma e^{\sigma}\sin\theta}{\theta} + 1 - e^{\sigma}\cos\theta)\mathbf{A} + \frac{1}{\theta^2+\sigma^2}(\sigma e^{\theta}\frac{1-\cos\theta}{\theta^2} - e^{\sigma}\frac{\sin\theta}{\theta} + \frac{e^{\sigma}-1}{\sigma})\mathbf{A}^2. \tag{5.11}$$

We emphasize that the proposed method does not fail for the degenerate cases such as the example case of Figure 5.5. In this example, each feature correspondence in Figure 5.5 can be represented by a similarity transformation matrix T_i, where $i \in \{1, 2, 3\}$. Since

$$T_1 = \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \neq T_2 = \begin{pmatrix} s_2 & 0 & \frac{x_1}{2} \\ 0 & s_2 & \frac{y_1}{2} \\ 0 & 0 & 1 \end{pmatrix}, \tag{5.12}$$

the matching of the first and the second features will not incorrectly contribute to the same bin in parameter space.

Although the proposed framework is presented for similarity invariant $n - th$ order visual features, the method is extendable to any geometrical transformation. For instance, the only change needed to adapt the algorithm to affine invariant kernel of $n-th$ order features is to define an affine transformation between every pair of spatio-temporal volumes of data similar to Equation 5.6. It is evident that in case of affine transformation the Rodrigues' formula does (Equation 5.9) not hold and therefore the computation of matrix logarithm becomes costly.

5.4 The Choreography Dataset

We collected a dataset of eight group dance styles from YouTube website from user upload videos. Currently, the videos for four classes are prepared properly by converting to a fixed frame rate of $30fps$ and dividing into at least eight short clips where each clips is dominated by a single dance routine. The statistics of the dataset is provided in Table 5.2. Figure 5.7 illustrates example frames from the dataset. Each category of the dataset is dedicated to a specific dance style. Different categories may share some similar dance routines but the collection of dance routines in one class is different from other classes.

5.4.1 Classification Strategy

In this section, we provide the outline of the classifier for classification of dance styles in the provided dataset. Figure 5.8 illustrates the block diagram of the proposed method for classification. In this setup, we use the kernel computation for visual phrases to train a SVM classifier to learn the class labels (dance styles) of a video clip. Provided the video clips of an input video example, the SVM classifier is capable of assigning a label to each clip separately. We recognize the label of

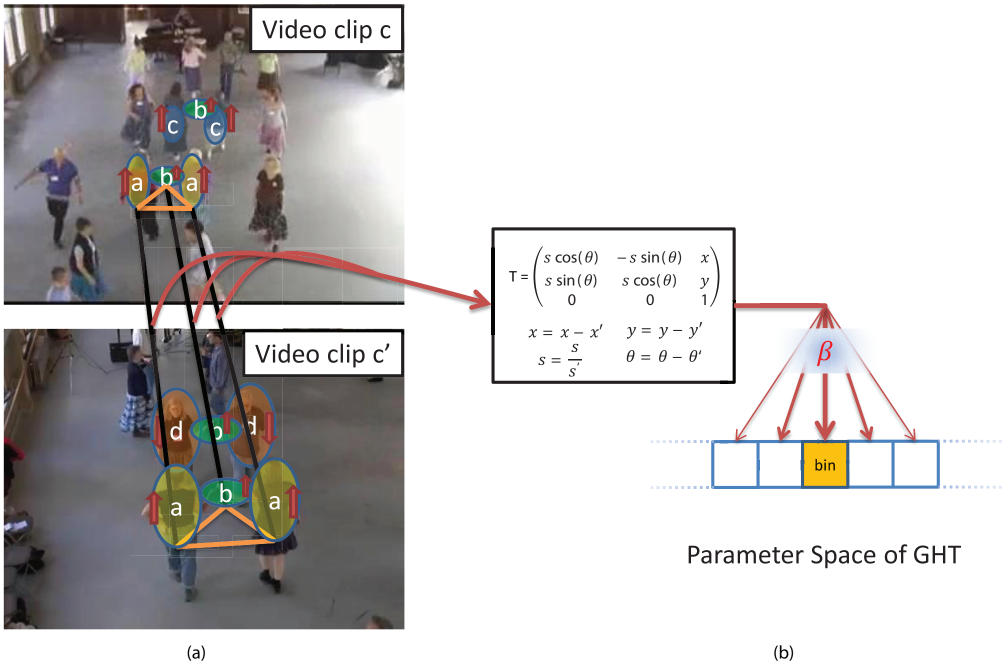

Video clip c
c
b
c
a
b
a
Video clip c'
d
b
d
a
b
a
$T = \begin{pmatrix} s\cos(\theta) & -s\sin(\theta) & x \\ s\sin(\theta) & s\cos(\theta) & y \\ 0 & 0 & 1 \end{pmatrix}$
$x = x - x'$
$y = y - y'$
$s = \frac{s}{s'}$
$\theta = \theta - \theta'$
β
bin
Parameter Space of GHT
(a)
(b)

Figure 5.6: Illustration of the algorithm to count the co-occurring visual phrases which are in certain geometrical layout in two video clips using a weighted Generalized Hough Transform. (a) Two frames of two different video clips are illustrated. Ellipses illustrate the visual words and the color and the labels refer to the word index. The size and orientation of visual words are represented by the size of the ellipse and the orientation of the major axis. The yellow triangle indicates the spatial relationship between three words. (b) Parameter space of GHT that considers translation, scale, and rotation between visual words. Each word correspondence contributes a weighted vote (β) to all bins in the parameter space. A group of words in the same spatial relationship contributes the most to the same bin. The similarity transformation is formed from the corresponding translation, scale, and orientation of the features to computes the voting weights.

the video example as the label with majority of the votes from the containing clips of that video. In this dataset, we provided the video clips for all of the video examples through manual annotation.

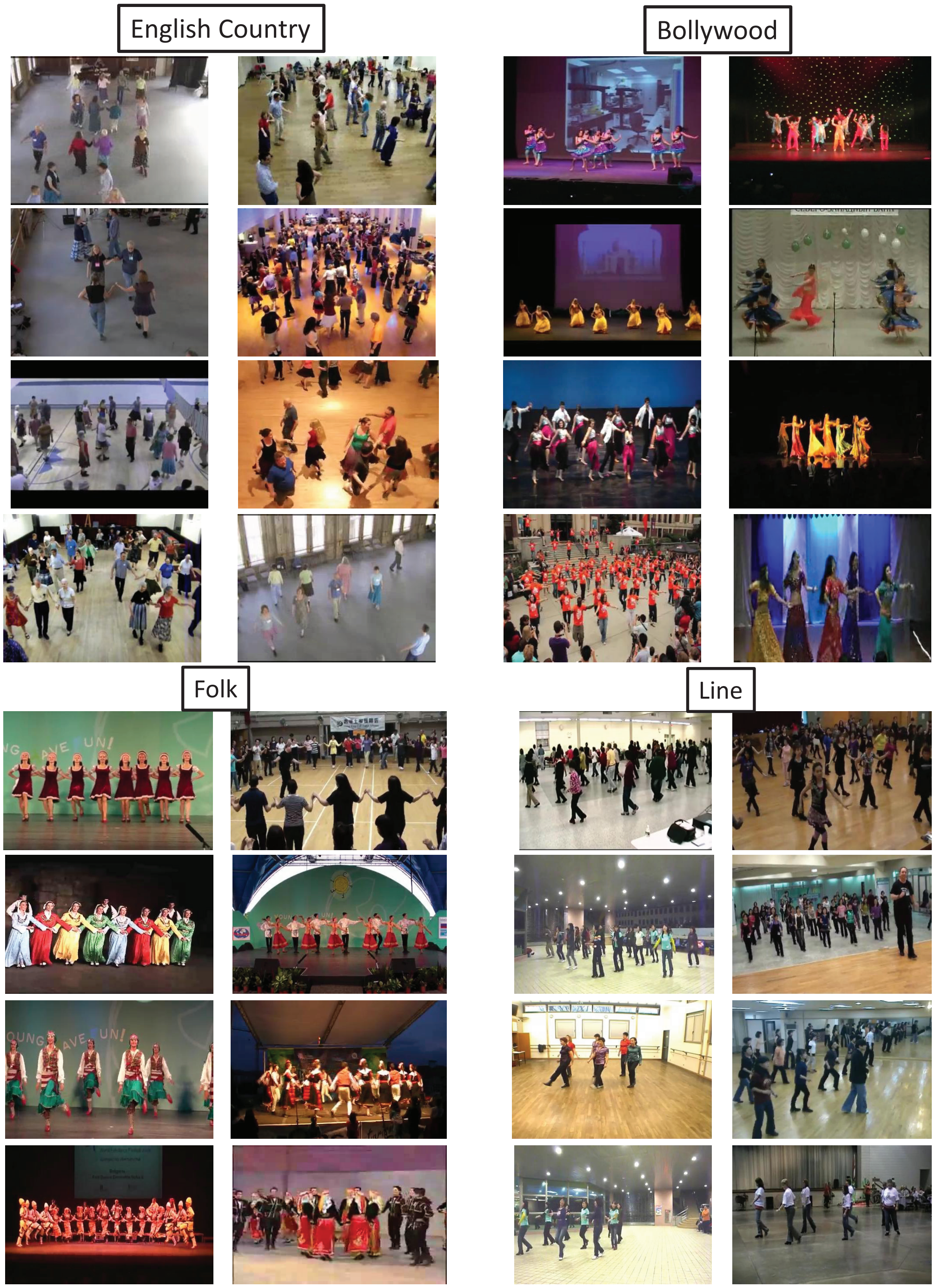


Figure 5.7: Sample frames of the four categories of group dances in the dataset.

Style	Examples	Routine Clips	Total # Frames
Ballet	12	116	17886
Bollywood	12	105	13529
English Country	12	138	16395
Folk	12	103	12600
Hiphop	12	99	14473
Kick	12	104	15073
Line	12	103	13017
Swing	12	107	15454

Table 5.2: The statistics of the collected dataset for group dances.

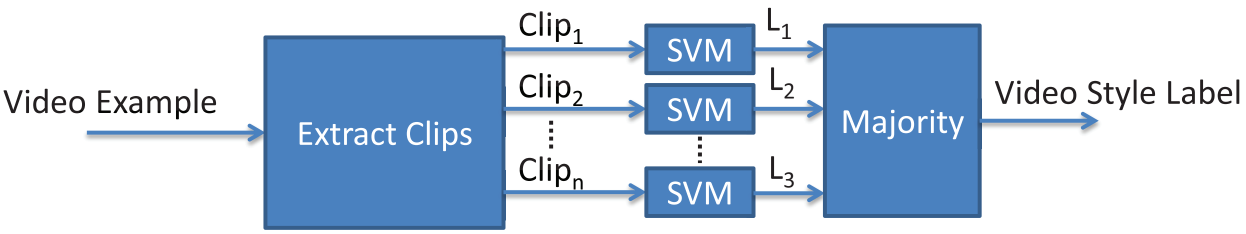


Figure 5.8: The classification block diagram for dance style recognition.

5.5 Experiments and Results

In this section, we describe the setup and the experiments to evaluate the proposed method. In these experiments, we used a subset of the collected dataset that contains 4 dance style categories of Bollywood, English Country, Folk, and Line which contains total of 321 video clips. The distribution of the number of clips in the subset is provided in Table 5.3. To demonstrate the effects of the proposed method, we extracted features as spatio-temporal cuboids of image intensity

Dance Style	# Examples	# Clips
Bollywood	8	73
English Country	8	106
Folk	8	72
Line	8	71

Table 5.3: Subset dataset (Bollywood, English Country, Folk, and Line).

Parameter	value	Role
σ	2	Spatial Scale
τ	1.5	Temporal scale
n	11	# of rows
m	11	# of columns
t	20	# of time slices

Table 5.4: Parameters of the feature extraction in [19].

using the method of [19] over the video clips. The cuboids have the same dimensions in space and time, and we ignore the temporal ordering between cuboids. During feature extraction, the center location of the the cuboids are assigned to their (x, y) and the average orientation of spatial gradients is assigned as the orientation (θ) of the feature by considering the spatial gradients from all of the time slices of a cuboid. In this experiment, the scale of the feature is considered the same ($s = 1$) for all of the the features. Table 5.4 summarizes the selected parameters for feature extraction using the method of [19].

The bag of words (BoW) representation is considered as the baseline method. For this purpose, we generate a codebook of size $k = 200$ visual words from the pool of spatio-temporal features detected over all video clips in the subset dataset using k-means clustering. The spatio-temporal features are mapped into a lower dimensional space ($dim = 200$) using PCA prior to applying k-means algorithm.

We implement several experiments to evaluate the effect of each part of the extensions proposed over the recent work in [77] which we refer here as *BoVP*. In BoVP method, higher order features of two video clips are matched invariant to translation, and $q = 441 = 21^2$ bins are used to quantize the parameter space. In the first set of experiments, we study the effect of the weighted Generalized Hough Transform using the weighting function in Equation 5.7 on BoVP method. We refer to this experiment by *Weighted BoVP* which involves weighted votes and transformation invariance of higher order features. In the second set of experiments, we use the weighted votes, invariance to rotation, and invariance to translation using the metric described in Equation 5.8, since the features contain the same scale factor. We refer to this experiment as *Weighted BoVP+*θ. In both *Weighted BoVP* methods, we used $q = 9 = 3^2$, and in *Weighted BoVP+*θ we use $q = 27 = 3^3$ bins.

Through the use of visual phrases, we are able to discriminate similar video clips through distinguishing groups of visual phrases in certain spatial layout. We verify this in Figure 5.9 by comparing the results of histogram intersection of visual words and visual phrases for the 321 video clips in the subset dataset. In these experiment, the visual phrases and histogram intersection is performed through use of the method in [77] which is digested in Section 5.3.2. In this experiment, the visual phrases are matched with invariance to translation, and Generalized Hough Transform with hard assignment is used to compute the histogram intersection in an effective manner. As illustrated in Figure 5.9, by including the higher order visual features in the matching process, the similarity map between video clips in the dataset converges toward a block diagonal matrix which is an expected behavior for discriminative features. This provides insight to understand better the discriminative power of the visual phrases in contrast to visual words.

Figure 5.10 and Table 5.5 summarize the recognition rates obtained by the proposed method and juxtapose them with the results from the state of the art method in [77] over the subset dataset. The recognition rate is obtained through leave-one-out cross-validation by excluding all the clips of one example video from every category. The examples are picked once randomly for 300 trials and the reported recognition rate is the average of these trials. In this experiment, we evaluated the methods by incrementally increasing the order of the visual phrases from 1 to 4, and observed the effect of higher order visual words. As Figure 5.10 indicates, the highest classification rate of all three methods is achieved when a higher order visual phrases is used. That is the case of 2nd order in BoVP method, 2nd order in Weighted BoVP method, and 3rd order in Weighted BoVP+θ method.

In addition, we observed that we obtained consistent improvement in the result using the proposed Weighted BoVP instead of hard assignments in BoVP over different orders of visual phrases. BoVP for visual phrases order 1 reduces to bag of words, however, Weighted BoVP and Weighted BoVP+θ do not. Therefore, the results are more appealing when we consider the improvement gained through the weighted methods for the 1st order visual phrases.

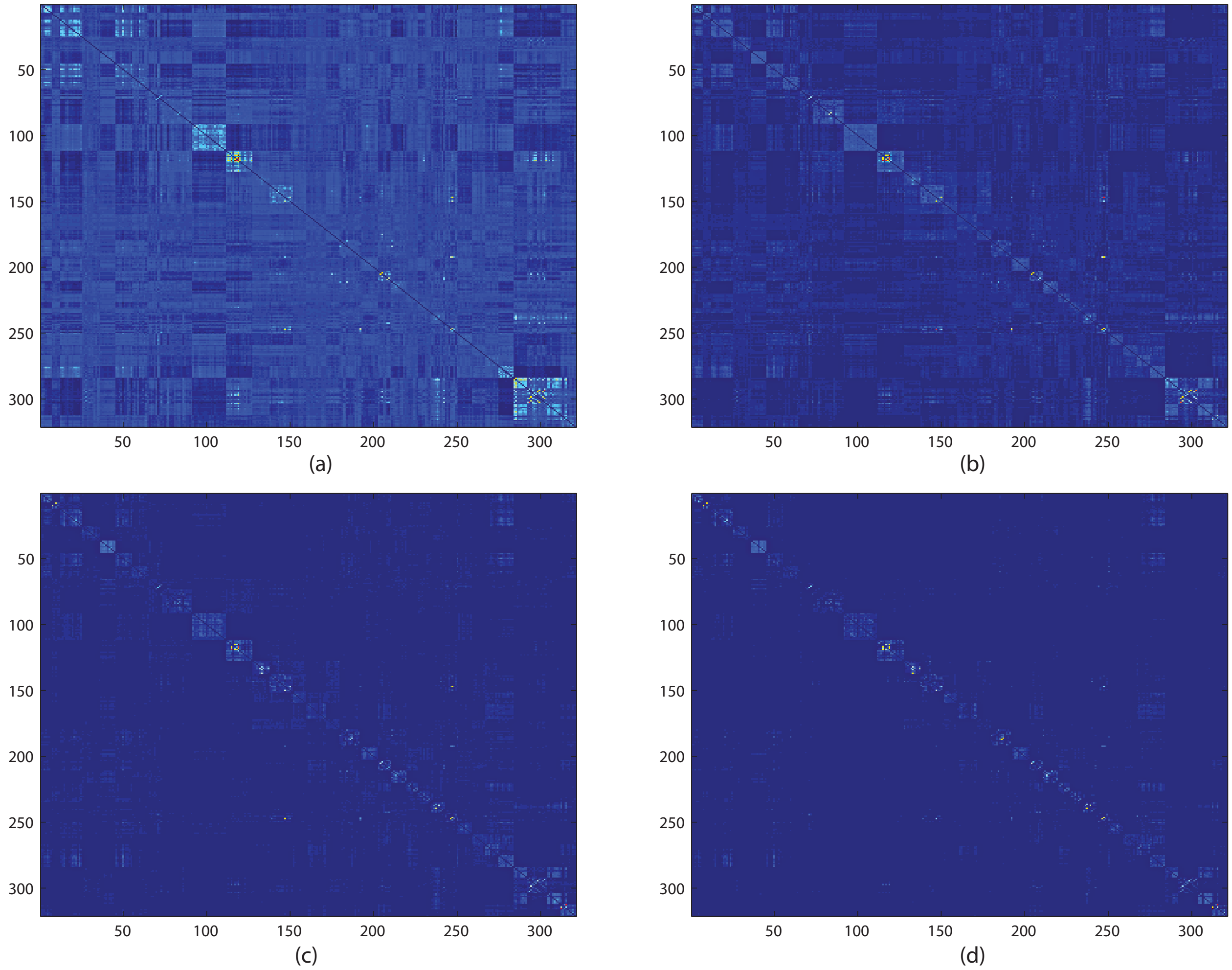


Figure 5.9: The similarity map between video clips in the subset dataset (321 video clips) using histogram intersection of visual word or phrases. Blue = low similarity, Red = high similarity. (a) Visual words (visual phrases of order 1), (b)visual phrases of order 2, (c) visual phrases of order 3, (d) visual phrases of order 4

The use of feature orientation improves the performance almost all orders of visual phrases, and the best performance is obtained through the use of Weight BoVP that includes up to 3rd order visual phrases that are invariant to rotation and translation. In particular, the orientation has improved the performance more than 10% (Table 5.5) when we considered higher order features.

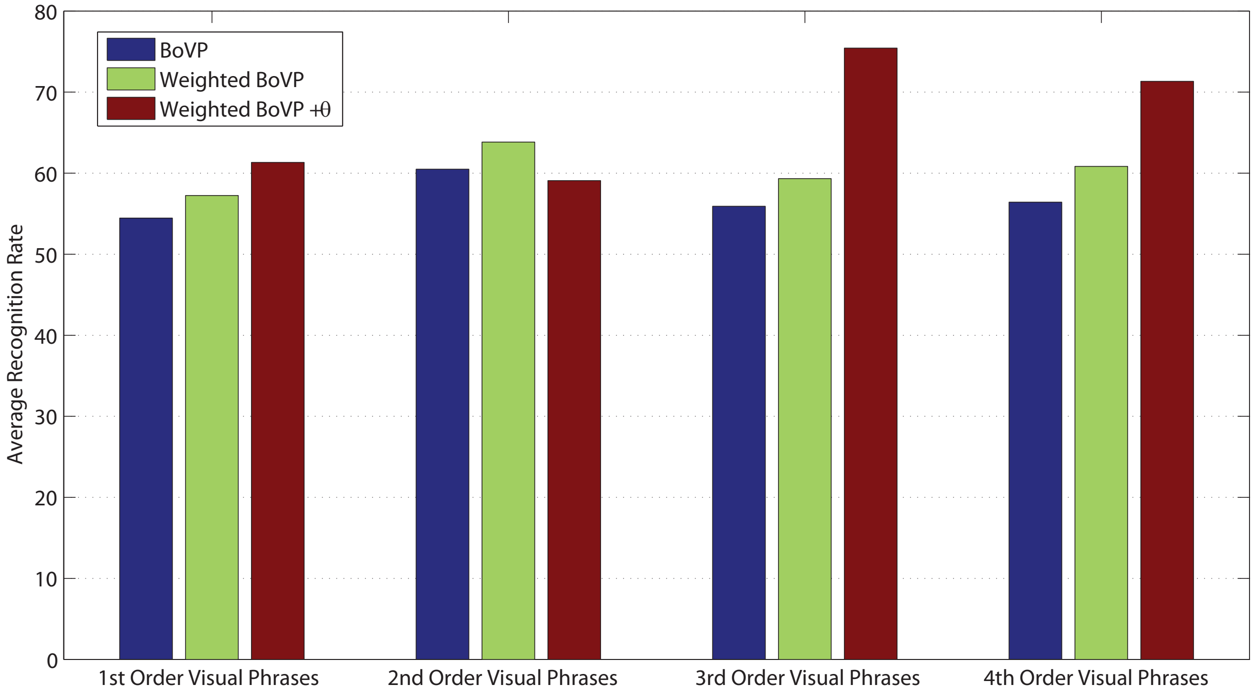
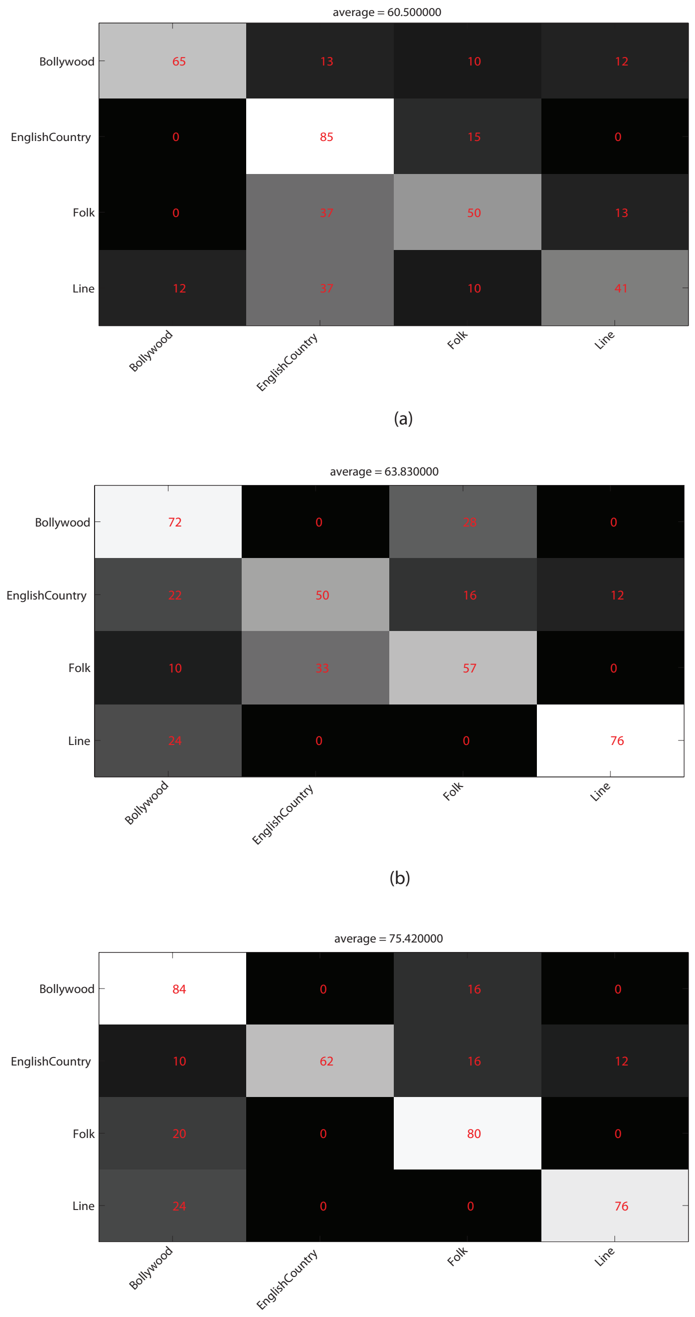


Figure 5.10: The average recognition rate (%) over four classes of the subset dataset using the proposed method in comparison to the state of the art [77] by considering different orders of visual phrases.

This indicates that the rotation invariance is more effective when we consider higher order features, or in other words, more complex layouts.

For better understanding the effects of the proposed methods in improving the classification performance, we provide the comparison of the confusion tables of BoVP, Weighted BoVP, and Weighted BoVP+θ in Figure 5.11. In this figure, the recognition rate and the confusion between each pair of classes is provided in percentage. In addition, the white color corresponds to the highest correlation and black indicates the lowest correlation.



(a)

(b)

(c)

Figure 5.11: The comparison of confusion tables for classification of 4 classes in the subset dataset, (a) BoVP, (b) Weighted BoVP, and (c) Weighted BoVP+θ.

Kernel Order	BoVP	Weighted BoVP	BoVP+θ
1	54.45	57.25	61.33
2	**60.5**	**63.83**	59.08
3	55.92	59.33	**75.43**
4	56.42	60.83	71.33

Table 5.5: Average recognition rate (%) on 4 class classification (Bollywood, English Country, Folk, and Line).

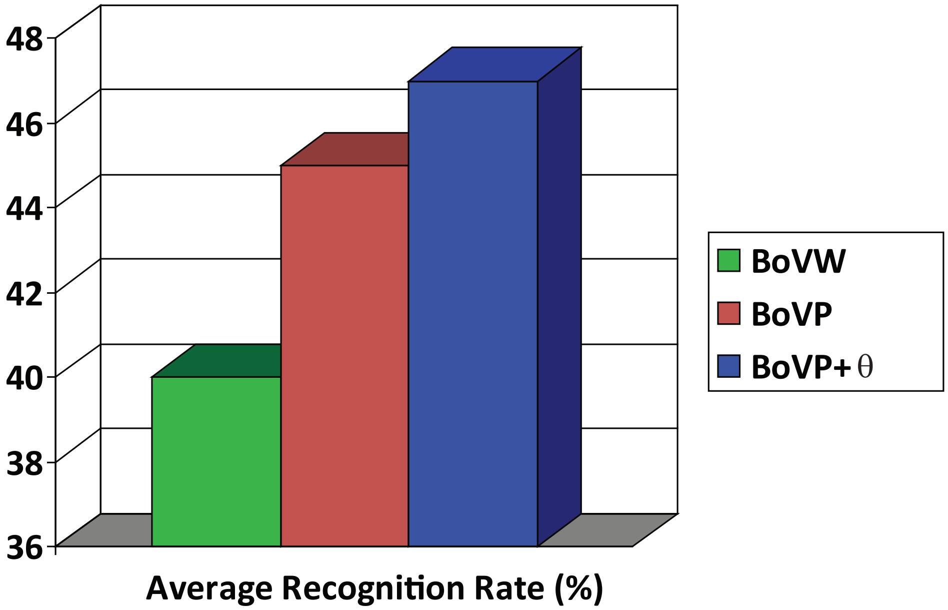


Figure 5.12: The average recognition rate over the full dataset.

Finally, we compared our proposed method to BoVP method over the full set. Figure 5.12 illustrates the same pattern of improving performance using Weighted BoVP and Weighted BoVP+θ methods over the full dataset. Figure 5.13 illustrates the confusion table of the our best results on

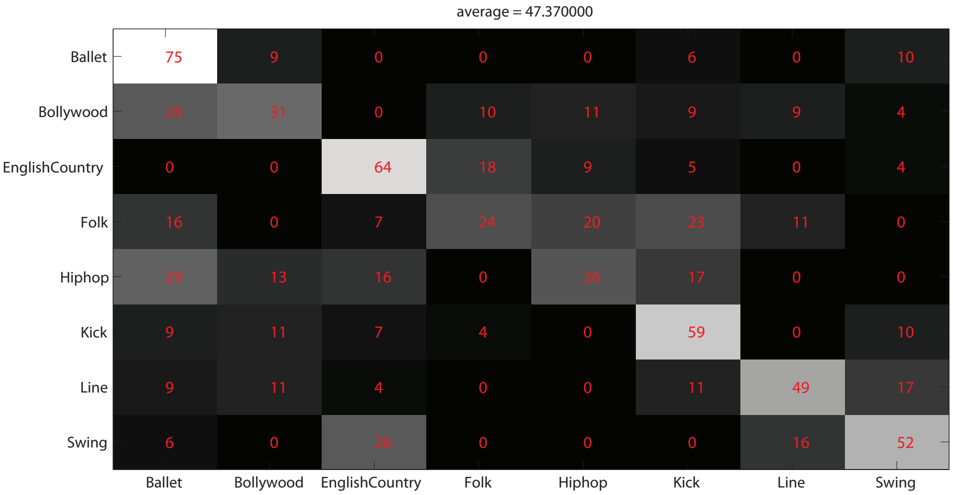


Figure 5.13: The confusion tables for classification of 8 classes in the subset dataset using Weighted BoVP+θ.

the full dataset using Weighted BoVP+θ method. Based on our observation which is in accordance to common knowledge of dances, Folk and Hip-hop dance styles are less structured dances with large variety of movement whereas Ballet and English Country dances are more structured and contain less variation in motion patterns. The results in Figure 5.13 provide an empirical support for our understanding of the difference between dance style. Where the structured dance styles exhibit less confusion compared to other dance styles.

CHAPTER 6: SUMMARY

In this work, we have explored the problem of human activities recognition in crowds through modeling different manifestations of cooperation within objects. At the first level of manifestations, objects in the crowd cooperate as a single entity and we modeled that by adapting Social Force Model for dense trajectories. Although it is successful in visual recognition of events in a crowded scene, this visual representation suffers from a major drawback that does not identify the different behaviors and entities in a scene. We proposed a method based on the concepts from fluid mechanics, Streaklines and Flow Potentials, to address these limitations and model crowds as mixtures of regions of different behaviors through motion segmentation. We further, recognized the behavior of the regions as lanes or convergent/divergent. Finally, to model the inter-relation of multiple entities in a scene, we extended a method of bag of visual phrases (BoVP) to capture the inter-relation of motion patterns within a scene. We collected a dataset of group choreographies to examine the proposed method. We summarize the major contributions in the following section.

6.1 Summary of Contribution

- Crowd event detection using Social Force Model
  - A method to adapt Social Force Model to video of crowds,

  - An algorithm for abnormal behavior detection based on the estimated interaction forces within a crowd.

- Behavior recognition using two new motion representations

  - Introducing Streaklines to computer vision community for analysis of dynamic flow in a crowded scene,
  - A novel algorithm to perform motion segmentation in scenes with dynamic flow,
  - A new algorithm recognizing divergent/convergent regions within a crowd using flow potentials,
  - A novel algorithm for detecting lanes based on motion segmentation and flow potentials.

- Behavior recognition through modeling the inter-relation of visual features

  - Adapting the bag of visual phrases to videos,
  - Extending the bag of visual phrases to include complete invariance to similarity transformation,
  - Introducing an efficient algorithm for soft assignments in Generalized Hough Transform for computation of histogram intersection of bag of visual phrases,
  - Introducing a new dataset to computer vision community for group behavior recognition: Group Choreography Dataset.

6.2 Future Work

The approaches proposed in this work have certain limitations and can be improved in many ways. We describe some attempts in the following subsections.

6.2.1 Limitation: Personal Desire Velocity in Social Force Model

In the Chapter 3, we followed the hypothesis that in a dense enough crowd, people tend to move with the group velocity and therefore, the the spatio-temporal average of the optical flow is a good estimation of the actual velocity of pedestrians. In this setup, the instantaneous optical flow is a good estimation of the desired velocity of a person. However, this is limiting the application of the approach to the cases where the desired velocity of the person is similar to the exact personal desire velocity. This limitation prevents the algorithm to estimate good quality interaction forces in scenes where there are many different points of interest or where the points of interest are moving and have a dynamic. To solve this problem, the algorithm should be able to make a educated guess regarding the desired location for each pedestrian based on the knowledge of the scene such as entry/exit points or the context.

6.2.2 Limitation: Streakline and Motion

The streakline representation is more helpful where the motion in the scene is temporally dense. For example in a traffic scene where the vehicles are moving densely, the streaklines are capable of well representing the flow. However, when the appearance of the vehicles are sparse in time, the streaklines that are computed directly from optical flow are not more informative than other representations of flow. To apply streaklines to other applications which involve sparse motion in time, we need an intermediate and dense representation of motion that is compiled from observing the scene for a duration of time.

6.2.3 Limitation: Potential and Global Motion

As Helmholtz theorem indicates, the decomposition of a flow field into two potential fields is only valid in absence of a global motion (Laminar flow). Therefore, the proposed method in Chapter 4 for behavior recognition is limited to cases with no camera motion. In case of a moving camera, the algorithm should estimate and cancel the global motion prior to any computation.

6.2.4 Limitation: Group Behaviors

In Chapter 5, the method for group behavior recognition does not distinguish between informative and uninformative visual words in learning a dance style. In addition, the method treats all the short clips in a video that contain dance routines equally without weighting them according to the relevance of the contained visual words. Therefore, the proposed approach is hardly scalable to other datasets or real videos. In order to mitigate this problem, the behavior learning model should change and methods like Multiple-instance Learning [8] are suggested.

6.2.5 Improvement: Interaction Forces within Groups

The algorithm we proposed using Social Force model for estimation of interaction forces between object in a scene is independent of the notion of groups. This method is effective, but it can fail in many ways in scenes involving different groups. A natural extension to this method is to incorporate a method to identify groups and compute the interaction between groups. The basic idea is to perform motion segmentation prior to estimating the interaction forces. This improves the estimation as it automatically eliminates the computation of interaction forces within groups.

6.2.6 Improvement: Streakline Representation

In this work, we introduced streaklines as a new representation of motion that is most suited for dynamically changing flows. This representation is applied to model the behavior of people in crowded scenes. However, this rich representation is general and can readily be applied to other computer vision problems. The problems that can benefit from this representation are those where their automatic visual perception is applied for detection of change in a dynamics or series of patterns. This clearly indicates applications in visual diagnostics in medical imaging.

6.2.7 Improvement: Flow Potentials

The flow potential representation is the second novel representation of flow in this work which is applicable to different problems in computer vision and computer graphics. In computer vision, the flow potentials capture a compact representation of the motion in a region. The peaks and valleys of these scalar fields are compressed representation of the motion. A graph representation of the relative peaks and valleys will represent the motion pattern of scene which can be used for visual behavior recognition. In computer graphics, a realistic crowd motion with a natural transition between different phases of behavior can be produced by modeling the changes in the location and magnitudes of the peaks and valleys of a flow field.

6.2.8 Bag of Visual Phrases using weighted Generalized Hough Transform

In this work, we proposed an efficient method to capture similarity invariant bag of visual phrases to model the inter-relation of motion pattern in group behaviors. However, this is a general concept and can readily be applied to other problems involving bag of phrases such as object recognition or scene recognition. The proposed method is directly applicable to image and video retrieval problems.

LIST OF REFERENCES

[1] History of stampedes, http://thoughtcatalog.com/2010/a-history-of-human-stampedes/.

[2] Unusual crowd activity dataset of University of Minnesota, available from http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi.

[3] Stampede near india shrine kills 100. *New York Times*, January 15 2011.

[4] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, to appear.

[5] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, 2007.

[6] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. *ECCV*, 2008.

[7] E. L. Andrade, S. Blunsden, and R. B. Fisher. Modelling crowd scenes for event detection. In *ICPR*, 2006.

[8] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*. MIT Press, 2003.

[9] G. Antonini, S. V. Martinez, M. Bierlaire, and J. P. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *IJCV*, 69(2):159–180, 2006.

[10] D. Bauer, S. Seer, and N. Brändle. Macroscopic pedestrian flow simulation for designing crowd control measures in public transport after special events. In *SCSC*, 2007.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[12] G.J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, 2006.

[13] Antoni B. Chan and Nuno Vasconcelos. Mixtures of dynamic textures. In *ICCV*, 2005.

[14] R. Chellappa. Machine recognition of human activities: A survey. *T-CSVT*, 18(11):1473–1488, September 2008.

[15] S.J. Cohen and Dance Perspectives Foundation. *International encyclopedia of dance: a project of Dance Perspectives Foundation, Inc*. Oxford University Press, 2004.

[16] T. Corpetti, E. Memin, and P. Perez. Extraction of singular points from dense motion fields: An analytic approach. *Journal of Mathematical Imaging and Vision (JMIV)*, 2007.

[17] N. Courty and T. Corpetti. Crowd motion capture. *Journal of Computer Animation and Virtual Worlds (JCAVW)*, 18(4-5):361–370, 2007.

[18] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

[19] P. Dollar. Behavior recognition via sparse spatio-temporal features. In *PETS*, 2005.

[20] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multi-scale, deformable part model. In *CVPR*, 2008.

[21] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 2007.

[22] P. T. Fletcher, C. Lu, and S. Joshi. Statistics of shape via principal component analysis on lie groups. In *CVPR*, 2003.

[23] W. Forstner and B. Moonen. A metric for covariance matrices. *Technical report, Dept. of Geodesy and. Geoinformatics, Stuttgart University*, 1999.

[24] C. Garate, P. Bilinski, and F. Bremond. Crowd event recognition using hog tracker. In *PETS*, 2009.

[25] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *ECCV*, 2008.

[26] S. Gu, Y. Zheng, and Carlo Tomasi. Critical nets and beta-stable features for image matching. In *ECCV*, 2010.

[27] F. R. Hama. Streaklines in a perturbed shear flow. *Physics Fluids*, 5:644–650, 1962.

[28] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, pages 487–490, 2000.

[29] D. Helbing, A. Johansson, and H. Z. Al-Abideen. The dynamics of crowd disasters: An empirical study. *Physical Review E*, 75:046109, 2007.

[30] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51:4282, 1995.

[31] J. Helman and L. Hesselink. Visualizing vector field topology in fluid flows. *CGA*, 11(3):36–46, 1991.

[32] R. L. Hughes. A continuum theory for the flow of pedestrians. *Transportation Research Part B: Methodological*, 36(6):507–535, July 2002.

[33] R. L. Hughes. The flow of human crowds. *Annual Review of Fluid Mechanics*, 35:169–182, 2003.

[34] A. Johansson, D. Helbing, and P. K. Shukla. Specification of the social force pedestrian model by evolutionary adjustment to video tracking data. *Advances in Complex Systems (ACS)*, 10(2):271–288, December 2007.

[35] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. In *BMVC*, 1995.

[36] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007.

[37] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009.

[38] T. I. Lakoba, D. J. Kaup, and N. M. Finkelstein. Modifications of the helbing-molnar-farkas-vicsek social force model for pedestrian evolution. *Simulation*, 81(5):339–352, 2005.

[39] L.D. Landau and E.M. Lifshitz. *Advanced Mechanics of Fluids*. Wiley, New York, 1959.

[40] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[41] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 2008.

[42] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Eurographics*, volume 26, 2007.

[43] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos ”in the wild”. In *CVPR*, 2009.

[44] J. Liu and M. Shah. Learning human action via information maximization. In *CVPR*, 2008.

[45] M. Luber, J.A. Stork, G.D. Tipaldi, and K.O. Arras. People tracking with human motion predictions from social forces. In *ICRA*, 2010.

[46] J. S. Marques, P. M. Jorge, A. J. Abrantes, and J. M. Lemos. Tracking groups of pedestrians in video sequences. In *Proc. CVPRW*, 2003.

[47] R. Mehran, A. Oyama, and M. Shah. Abnormal behavior detection using social force model. In *CVPR*, 2009.

[48] M. Moakher. Means and averaging in the group of rotations. *SIAM J. on Matrix Analysis and Applications (SIMAX)*, 24:1–16, January 2002.

[49] J. C. Niebles, H. Wang, , and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[50] X. Pan, C. S. Han, K. Dauber, and K. H. Law. Human and social behavior in computational modeling and analysis of egress. *Automation in Construction*, 15(4):448–461, 2006.

[51] V. Parameswaran and R. Chellappa. View invariance for human action recognition. In *IJCV*, 2006.

[52] N. Pelechano, J. M. Allbeck, and N. I. Badler. Controlling individual agents in high-density crowd simulation. In *SCA*, 2007.

[53] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.

[54] P. Reisman, , O. Mano, S. Avidan, and A. Shashua. Crowd detection in video sequences. In *Intelligent Vehicles Symposium (IV)*, 2004.

[55] I. Saleemi, L. Hartung, and M. Shah. Scene understanding by statistical modeling of motion patterns. In *CVPR*, 2010.

[56] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. In *CVPR*, 2006.

[57] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80:72–91, 2008.

[58] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[59] E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. In *CVPR*, 2007.

[60] W. Shao and D. Terzopoulos. Autonomous pedestrians. *Graphical Models*, 69(5-6):246–274, 2007.

[61] A. Sud, R. Gayle, E. Andersen, S. Guy, M. Lin, and D. Manocha. Real-time navigation of independent agents using adaptive roadmaps. In *VRST*, 2007.

[62] W. Sultani and J. Y. Choi. Abnormal traffic detection using intelligent driver model. In *ICPR*, 2010.

[63] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.

[64] A. Treuille, S. Cooper, and Z. Popović. Continuum crowds. *ACM Transactions on Graphics (TOG)*, 25(3):1160–1168, 2006.

[65] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu. Unified crowd segmentation. *ECCV*, 2008.

[66] O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3d motion estimation via mode finding on lie groups. In *ICCV*, 2007.

[67] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d examplars. In *ICCV*, 2007.

[68] V. Wijk and J. Jarke. Image based flow visualization. In *SIGGRAPH*, 2002.

[69] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantics and structural information. In *CVPR*, 2007.

[70] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *CVPR*, 2009.

[71] W. Yu and A. Johansson. Modeling crowd turbulence by many-particle simulations. *Physical Review E*, 76(4):046105, 2007.

[72] J. Yuan, Z. Liu, and Y. Wu. Discriminative sub volume searches for efficient action detection. In *CVPR*, 2009.

[73] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007.

[74] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, and L. Xu. Crowd analysis: a survey. *Machine Vision and Applications (MVA)*, 19(5-6), October 2008.

[75] Y. Zhang and T. Chen. Efficient kernels for identifying unbounded-order spatial features. In *CVPR*, 2009.

[76] Y. Zhang and T. Chen. Weakly supervised object recognition and localization with invariant high order features. In *BMVC*, 2010.

[77] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, 2011.