BEYOND ACTIONS: DISCRIMINATIVE MODELS FOR CONTEXTUAL GROUP ACTIVITIES

by

Tian Lan
B.Eng., Huazhong University of Science and Technology, China, 2008

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the School
of
Computing Science

© Tian Lan 2010
SIMON FRASER UNIVERSITY
Summer 2010

All rights reserved. However, in accordance with the Copyright Act of Canada, this work may be reproduced, without authorization, under the conditions for Fair Dealing. Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.

APPROVAL

Name:	Tian Lan
Degree:	Master of Science
Title of Thesis:	Beyond Actions: Discriminative Models for Contextual Group Activities
Examining Committee:	Dr. Ze-Nian Li Chair
	Dr. Greg Mori, Senior Supervisor
	Dr. Shahram Payandeh, Supervisor
	Dr. Mark Drew, SFU Examiner
Date Approved:	Aug. 12, 2010

SFU SIMON FRASER UNIVERSITY
LIBRARY

Declaration of
Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the "Institutional Repository" link of the SFU Library website <www.lib.sfu.ca> at: <http://ir.lib.sfu.ca/handle/1892/112>) and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, BC, Canada

Abstract

Human action recognition from realistic videos is a challenging problem in computer vision. Several intrinsic properties such as intra-class variations, background clutter and partial occlusion make it difficult to recognize individual person actions reliably.

In this dissertation, we go beyond recognizing individual person actions and focus on group activities instead. This motivates from the observation that human actions are rarely performed in isolation, the contextual information of what other people nearby are doing provides useful cues for understanding the high-level activities. We propose a discriminative model for recognizing group activities. Our model jointly captures the group activity, the individual person actions, and the interactions among them. Two new types of contextual information, *group-person interaction* and *person-person interaction*, are explored in a latent variable framework. In particular, we propose two different approaches to model the person-person interaction. One approach is to explore the structures of person-person interaction. Different from most of the previous latent structured models which assume a pre-defined structure for the hidden layer, e.g. a tree structure, we treat the structure of the hidden layer as a latent variable and implicitly infer it during learning and inference. The other approach explores the person-person interaction in feature level. We introduce a new feature representation called the *action context (AC) descriptor*. The AC descriptor encodes information about not only the action of an individual person in the video, but also the behaviour of other people nearby. Our experimental results demonstrate the benefit of using contextual information for disambiguating group activities.

Keywords: computer vision; group activity recognition; context; latent structured models

Acknowledgments

First I would like to especially thank my advisor Professor Greg Mori, for his support and guidence in the past one and a half years. Greg directed me into the fascinating realm of computer vision, and has spent great efforts in helping me through many obstacles during my thesis. From our many discussions, I have been inspired by his deep insights in various computer vision problems and his amazing ability to organize and explain detailed concepts in a simple way that everyone can understand. I have learnt from him not only to be a passionate researcher but also to be a man of integrity, patience and kindness. I consider myself lucky for having been his student.

Thanks to Professors Mark Drew and Shahram Payandeh for serving on my thesis commitee and their insightful comments on my thesis. Thanks to Professor Ze-Nian Li for chairing my thesis defense. Thanks to Professor Stephen Robinovitch and SFU Injury Prevention and Mobility Lab for collecting the nursing home data for my thesis and their constructive comments on my research.

Dr. Yang Wang and Weilong Yang are my main collaborators in research. I'm deeply grateful for their generous help when I first started computer vision and their helpful advices and suggestions during these years. I have greatly enjoyed our fruitful discussions in research and so much joy and laughter we shared in life.

Thanks to all my labmates at VML including (in alphabetical order) Bo Gao, Jiawei Huang, Zhi Feng Huang, Bahman Khanloo, Mohammad Norouzi, Mani Ranjbar, Ferdinand Stefanus, Arash Vahdat, Pengfei Yu, Ziming Zhang, and many others. Also a big thank you to all my friends at various stages of my life, your support, encouragment and humor have sustained me and made my life joyful.

Finally, my deepest gratitude goes to my parents for their love and support. It is only through their dedication and sacrifices that I could have come this far.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Vision-based activity recognition has broad applications in various aspects of people's life, such as surveillance, human-computer interaction, robot learning, sports and entertainment, etc. At the highest level, the goal is to enable computers to analyze and understand human behavior. The task is difficult, the appearance of human activities has tremendous variation due to background clutter, partial occlusion, scale and viewpoint change, etc. During the last few decades, researchers have developed a number of methods and achieved remarkable success. However, we still have a long way to go.

Automatic video surveillance is an important application of vision-based activity recognition. Thousands of hours of videos are being captured everyday by CCTV camera, web camera, surveillance camera, etc. However, the technology of automatic video analysis has failed to keep pace and all of the task is left to human security personnel currently. This endeavour requires the analysis of a large amount of video recordings, and this task is not well suited to humans, as it is labor-intensive and demanding to sift through all the data. Under this background, a system that can automatically recognize and annotate all the activities occur in a video is required. Several state-of-the-art benchmark datasets are trying to address this issue, a representative one is the TRECVid [37] event detection dataset. The objective of TRECVid event detection is to automatically detect the observable events in surveillance videos, which are captured from London Gatwick airport. Typical events in this dataset include running, embrace, pointing, etc. Note that this task only requires to locate the frames which contain the pre-defined events, instead of accurate spatial locations. Around 100 hour video dataset has been released for the development purpose, which consists of the videos from five surveillance cameras in the airport. The representative frame

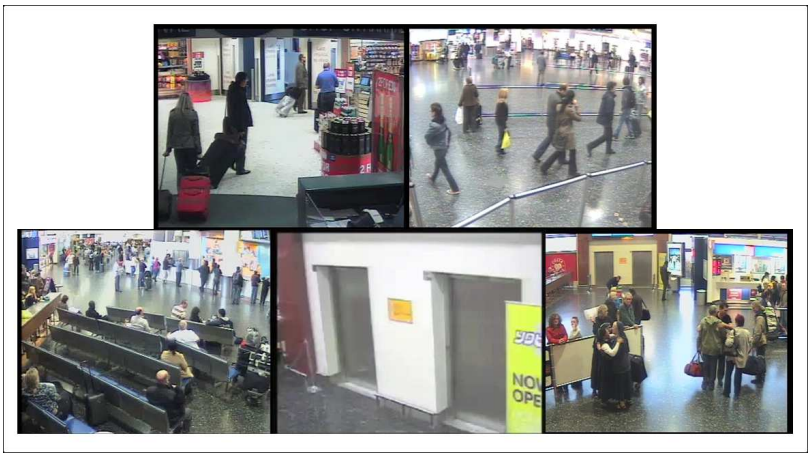


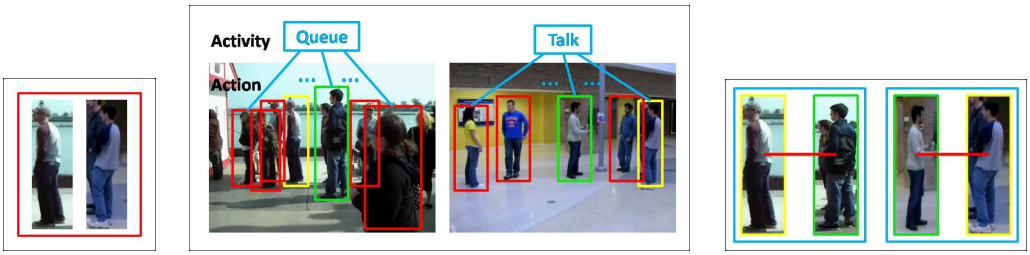
Figure 1.1: Representative Frames of TRECVID event detection dataset.

of each camera view are shown in Fig.1.1.

In many real-world applications, such as surveillance, reliably recognizing each individual's action using state-of-the-art techniques in computer vision is unachievable. One alternative is focusing on activity over a group of people instead. This motivates from the observation that human actions are rarely performed in isolation, the actions of individuals in a group can serve as context for each other. The goal of this dissertation is to explore the benefit of contextual information in group activity recognition in challenging real-world applications.

1.1 Background

Group activity recognition: Human activity understanding is of great scientific interest in the computer vision community. *Group activity recognition* is an important component of automatic human activity understanding. The goal of group activity recognition is to classify a video sequence, clip, or individual frames into several pre-defined categories according to activities performed by groups of people in the video. Most of the work in human activity understanding only focuses on single-person action recognition. In this dissertation, we argue that actions of individual humans often cannot be inferred alone. Look at the two persons in Fig. 1.2(a), can you tell they are doing two different actions? Once the entire contexts of these two images are revealed (Fig. 1.2(b)) and we observe the interaction of the person with other persons in the group, it is immediately clear that the first person is



(a) (b) (c)

Figure 1.2: Role of context in group activities. It is often hard to distinguish actions from each individual person alone (a). However, if we look at the whole scene (b), we can easily recognize the activity of the group and the action of each individual. In this dissertation, we operationalize on this intuition and introduce a model for recognizing group activities by jointly consider the group activity, the action of each individual, and the interaction among certain pairs of individual actions (c).

queuing, while the second person is talking. We instead focus on developing methods for recognizing group activities by modeling the collective behaviors of individuals in the group.

Before we proceed, we first clarify some terminology used throughout the rest of the dissertation. We use *action* to denote a simple, atomic movement performed by a single person. We use *activity* to refer to a more complex scenario that involves a group of people. Consider the examples in Fig. 1.2(b), each frame describes a group activity: queuing and talking, while each person in a frame performs a lower level action: talking and facing right, talking and facing left, etc.

Our goal is to perform group activity recognition in challenging real-world conditions, e.g. surveillance video data. Consider the video frames shown in Fig. 1.3. These are example frames from a nursing home surveillance video in which we would like to recognize instances of activities of interest such as residents who fall. The intra-class variation in activity categories and relatively poor video quality typical of surveillance footage render this a challenging problem.

The Role of Context: As indicated by psychology experiments, *context* is critical in recognition for human visual system [3]. In computer vision, the use of context is also important for solving various recognition problems, especially in situations with poor viewing quality. This is because features are usually not reliable in such circumstances, thus analysis of individual objects alone can not yield reliable results.

Figure 1.3: Sample frames from a nursing home surveillance video. Our goal is to find instances of residents falling down.

In this dissertation, we focus on exploring the role of context in group activity recognition. Contextual information is extracted from the behaviour of all the people in a video frame, as indicated in Fig. 1.2. As mentioned earlier, activity recognition from surveillance video data is challenging. With this type of video footage many actions are ambiguous, as shown in Fig. 1.3. For example, falling down and sitting down are often confused – both can contain substantial downward motion and result in similarly shaped person silhouettes. A helpful cue that can be employed to disambiguate situations such as these is the context of what other people in the video are doing. Given visual cues of large downward motion, if we see other people coming to aid then it is more likely to be a fall than if we see other people sitting down.

Here we define two types of contextual information in group activities exploited in this dissertation. First, the activity of a group and the collective actions of all the individuals serve as context (we call it the *group-person interaction*) for each other, hence should be modeled jointly in a unified framework. As shown in Fig. 1.2, knowing the group activity (queuing or talking) helps disambiguate individual human actions which are otherwise hard to recognize. Similarly, knowing most of the persons in the scene are talking (whether facing right or left) allows us to infer the overall group activity (i.e. talking). Second, the action of an individual can also benefit from knowing the actions of other surrounding persons (which we call the *person-person interaction*). For example, consider Fig. 1.2(c). The fact that the first two persons are facing the same direction provides a strong cue that both of them are queuing. Similarly, the fact that the last two persons are facing each other indicates they are more likely to be talking.

1.2 Contributions

In this dissertation, we develop a discriminative model for recognizing group activities. We highlight the main contributions of our model.

- *Group activity:* most of the work in human activity understanding focuses on single-person action recognition. Instead, we present a model for group activities that dynamically decides on interactions among group members.
- *Group-person and person-person interaction:* although contextual information has been exploited for visual recognition problems, ours introduces two new types of contextual information that have not been explored before.
- *Context descriptor:* In terms of person-person interaction, one way is to model it in feature level, i.e. the feature descriptor for each person could reflect actions of both the focal person and context simultaneously. We propose a context descriptor encodes information about an individual person in a video, as well as other people nearby.
- *Structure learning:* The other way is to model the interaction in structure level. The person-person interaction poses a challenging problem for both learning and inference. If we naively consider the interaction between every pair of persons, the learning and inference turn out to be intractable. Ideally, we would like to consider only those person-person interactions that are strong. To this end, we propose a structure learning approach that automatically decide on whether the interaction of two persons should be considered. Our experimental results show that our structure learning significantly outperforms other alternatives.

1.3 Outline

The rest of the dissertation is organized as follows:

Chapter 2 provides an overview of previous work in both computer vision and machine learning areas that is most relevant to this dissertation. Relevant work in vision includes human action/activity recognition and context based visual recognition, work in learning includes discriminative latent models.

Chapter 3 proposes a discriminative model for group activity recognition. Our model jointly captures the group activity, the individual person actions, and the interactions among

them. Context has been extensively studied in scene and object recognition. In this work, we demonstrate that it is also useful for group activity recognition. We introduce two different ways to model the person-person interaction, one is on exploring the structures, and the other is based on a context feature descriptor.

Chapter 4 concludes this thesis and discusses future work.

Chapter 2

Previous work

In this chapter, we give a general overview of previous work that is related to this dissertation.

2.1 Human Action/Activity Recognition

In the last few decades, a lot of work has been done in recognizing human actions/activities from video sequences or still images. There is a huge literature in this area and we only focus on the closely related work in this review.

Datasets: Most existing work test their algorithms on standard benchmark datasets, like KTH [36] and Weizmann [4], which normally only involve a single actor performing certain actions in a controlled setting with small camera motion and clean background. Many work has achieved high recognition accuracy in these datasets. Some recent work tests their algorithms on more complicated datasets that are close to the real-world conditions. Action recognition "in the wild" is receiving a lot of attentions. Several representative datasets are Hollywood Human Actions (HOHA) dataset [23], Youtube Action dataset [24] and Collective Activity Dataset [6]. There is also some work focuses on action recognition from still images [41, 20, 15, 9], progress made in still images can be directly applied to videos. Some representative datasets are shown in Fig. 2.1.

Bag-of-words Representation: The feature descriptor in our work employs a *bag-of-words* style representation. Bag-of-words representations have been studied extensively in computer vision, particularly in object recognition. In action recognition, Wang and Mori [42] track individual people and model co-occurances of the actions in a single track

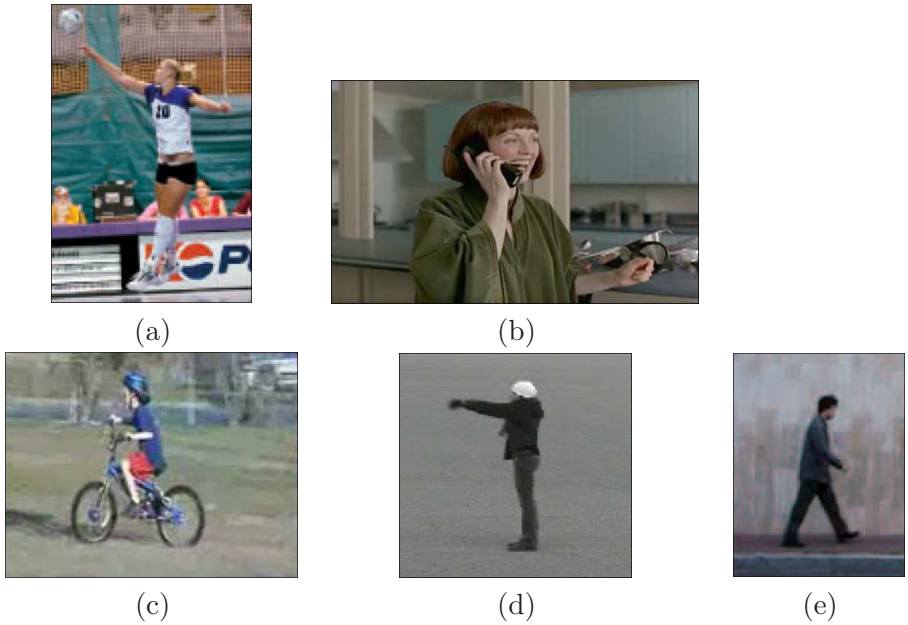


Figure 2.1: Example frames from several representative datasets. From (a)-(e) are: Human-Object Interaction [15], Hollywood [23], Youtube [24], KTH [36] and Weizmann [4] Dataset respectively.

with a mapping of frames to visual words. In contrast, the method we present here does not require tracking, which is challenging in our datasets, and models the actions of multiple people. Wang et al. [40] analyze far-field traffic video. Low-level atomic events are described by motion and position features, and hierarchical models are used to capture the co-occurances of these atomic events over video clips. We explicitly model the spatial context of an individual person, rather than treating the whole frame in a bag-of-words representation. Loy et al. [25] develop a structure learning algorithm to model temporal dependencies of actions across a camera network. Our model focuses on a lower level of detail, on the actions of an individual.

Group Activity Recognition: Most work on human activity recognition consider activities of one individual or between two individuals, such as hand-shaking, embracing and kissing. Relatively few work attempts to model the high-level group activities [31, 22, 50, 29, 21, 27, 35]. Most of the work on group activity focuses on a small range of activities with clear structural information using sequential models. For example, Vaswani et al. [31] models an activity using a polygon and its deformation over time. Each person in the group is treated as a point on the polygon. The model is applied to abnormality detection. Ivanov

and Bobick [21] divides the recognition problem into two levels. The first level detects the individual events with HMMs and the second level models the interactions among the events detected in the first level using a stochastic context-free grammar parsing mechanism. However, only sequential relations are considered in the parsing mechanism. Moore and Essa [29] extend the work in [21] for recognizing multitasked activities. The main limitation of this line of work is that the temporal ordering of the activities has to be strictly sequential. Thus, only a limited range of activities can be modeled by these approaches. More recently, Ryoo and Aggarwal [35] proposes a stochastic representation for group activities based on context-free grammar, which characterizes both spatial and temporal arrangements of group members. However, the representation of activities are encoded manually by human experts. Different from the above mentioned approaches, our work employs a latent variable framework that is able to capture the complex structures of various types of group activities, and the structures of group activities are learnt automatically.

2.2 Recognition with Context

Using context to aid visual recognition has received much attention recently. Some work clearly demonstrates that contextual information could improve the performance of weak local detectors [19, 39, 30, 5, 45, 51]. Most of the work on context is in scene and object recognition. For example, work has been done on exploiting contextual information between scenes and objects [30], objects and objects [8, 33, 13, 14], objects and so-called "stuff" (amorphous spatial extent, e.g. trees, sky) [18], etc.

Most of the previous work in human action recognition focuses on recognizing actions performed by a single person in a video (e.g. [4, 36]). In this setting, there has been work on exploiting contexts provided by scenes [26] or objects [17] to help action recognition. In still image action recognition, object-action context [9, 15, 47, 48] is a popular type of context used for human-object interaction. In this dissertation, we focus on another type of contextual information – the interactions between people. Modeling interactions between people and their role in action recognition has been explored by many researchers. For example, sophisticated models such as dynamic Bayesian networks [46] and AND-OR graphs [16] have been employed. Gupta et al. [16]'s representation based on AND-OR graphs allows for a flexible grammar of action relationships. The sophistication of these models leads to more challenging learning problems. Other representations are holistic in

nature. Zhong et al. [52] examine motion and shape features of entire video frames to detect unusual activities. Mehran et al. [28] build a "bag-of-forces" model of the movements of people in a video frame to detect abnormal crowd behaviour. The work in [6] is the closest to ours. In that work, pose-pose context is exploited by a new feature descriptor extracted from a person and its surrounding area.

2.3 Discriminative Latent Models

Our model is directly inspired by some recent work on learning discriminative models that allow the use of latent variables [1, 9, 32, 43, 49], particularly when the latent variables have complex structures. These models have been successfully applied in many applications in computer vision, e.g. object detection [12], action recognition [43], human-object interaction [9]. So far only applications where the structures of latent variables are fixed have been considered, i.e. a tree-structure in [12, 43]. However in our applications, the latent structures are not fixed and have to be inferred automatically.

Latent SVM

Among the discriminative latent models, latent SVM [12] is the closest to our model. This model is extended to solve multi-class classification problem, known as latent structural SVM [49] or max-margin hidden conditional random field [43]. In [12], latent SVM is proposed and applied to object detection. The model combines the flexibility of part-based approaches and global perspective of large-scale features in a unified framework. It is also an integration of discriminative structural models and max-margin learning principle.

In the latent SVM framework, the models are trained with partially labeled data. In [12], object locations are labeled while part locations are not labeled and treated as latent variables during training. The model for an object consists of an appearance model and a spatial model. An example of the model is shown in Fig. 2.2.

Let x be an example and z be the configurations of x. We assume that z takes the form of $z = (p_0, p_1, \ldots, p_K)$, where p_0 is location of the root (the whole object), $p_i (i = 1, \ldots, K)$ is the location of i-th part and K is the number of parts. We use an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the configuration z. In [12], G is defined as a star graph with a root plus a collection of parts connected to it, where a vertex $v_i \in \mathcal{V}$ corresponds to the location of the i-th part, and an edge corresponds to the displacement of the j-th part to the root.

An example x is scored with a function of the following form:

$$f_w(x) = \max_{z \in \mathcal{Z}(x)} w^\top \Psi(x, z) \tag{2.1}$$

where w is a vector of model parameters and z are latent variables. $\mathcal{Z}(x)$ is the set of all possible latent values for an example x. The model parameters w are simply the combination of two parts, $w = \{\alpha, \beta\}$ and $w^\top \Psi(x, z)$ is defined as:

$$w^\top \Psi(x, z) = \sum_{j=0}^{K} \alpha^\top \phi(x, p_j) + \sum_{j=1}^{K} \beta^\top \varphi(p_j) \tag{2.2}$$

where $\phi(x, p_j)$ denotes the appearance feature of the root when $j = 0$ and appearance feature of the j-th part if $j > 0$. $\varphi(p_j)$ denotes the spatial feature which represents the displacement of j-th part relative to the root.

Given a set of N training examples $\langle x^n, y^n \rangle (n = 1, 2, \ldots, N)$, where $y^i \in \{-1, 1\}$, the model parameter $\mathbf{w}$ is trained with the following formulation:

$$\min_{w, \xi \geq 0} \ \frac{1}{2}||w||^2 + C \sum_{n=1}^{N} \xi_n \tag{2.3a}$$

$$\text{s.t.} \quad y^n f_w(x^n) \geq 1 - \xi_n, \forall n \tag{2.3b}$$

Note that if the set of latent values $\mathcal{Z}(x)$ is fixed for each example x, then the formulation of latent SVM is the same as regular SVM. The latent SVM leads to a semi-convex optimization problem as defined in [12]: The optimization problem in Eq. 2.3 is convex for negative examples and non-convex for positive examples. A coordinate descent algorithm is introduced in [12] to compute a local optimum of Eq. 2.3:

1. Holding w, ξ fixed, optimize the latent variable z' for each positive example x_i:

$$z_i = \arg \max_{z' \in \mathcal{Z}(x_i)} w^\top \Psi(x_i, z') \tag{2.4}$$

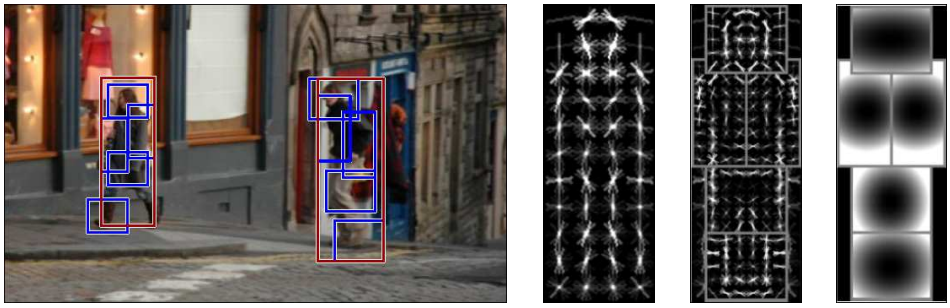
2. Holding z_i fixed, optimize w, ξ by solving the convex optimization problem:

$$\min_{w, \xi \geq 0} \ \frac{1}{2}||w||^2 + C \sum_{n=1}^{N} \xi_n \tag{2.5a}$$

$$\text{s.t.} \quad y^n f_w^p(x^n) \geq 1 - \xi_n, \forall n \tag{2.5b}$$

where $f_w^p(x)$ is defined as: $f_w^p(x) = \max_{z \in \mathcal{Z}^p} w^\top \Psi(x, z)$, $\mathcal{Z}^p$ is obtained by restricting the latent values for the positive examples according to z_i.

We will build on this model in our work.



(a) (b) (c) (d)

Figure 2.2: Examples of detections obtained by latent SVM model shown in (a). The model is defined by a root filter with lower resolution (b), several part filters with higher resolution (c) and a spatial model indicates the relative location of each part w.r.t. the root (d). The root filter and part filters are the learnt weights for the appearance features of root and parts respectively. The images are from [12]

Chapter 3

Group Activity Recognition with Context

As mentioned earlier, human activity recognition in real-world conditions is a very challenging computer vision problem. In order to reliably interpret the high-level human activity, it is likely that contextual information of the individual actions under the same scene will need to be explored.

In this work, two new types of contextual information, *group-person interaction* and *person-person interaction*, are explored in a latent variable framework. Central to our problem is how to model the person-person interaction, we develop two different approaches to solve this problem. One approach is to explore the structures of person-person interaction (Sec. 3.1). Different from most of the previous work in latent structured models which assume a predefined structure for the hidden layer, e.g. a tree structure, we treat the structure of the hidden layer as a latent variable and implicitly infer it during learning and inference. Intuitively speaking, the structure learning approach will automatically decide on whether the interaction of two persons should be considered. The other approach explores the person-person interaction in feature level (Sec. 3.2). We propose a context descriptor which encodes information about not only the action of an individual person in the video, but also the behaviour of other people nearby. This feature representation is inspired by the fact that the context of what other people are doing provides very useful cues for recognizing the actions of each individual.

We assume an image has been pre-processed so the persons in the image have been found.

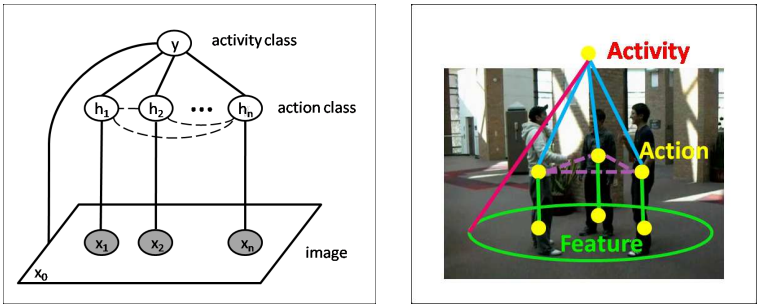
How to localize people in the video frames is task-specific, and it involves either human detection [12] or background subtraction. We will describe the details in the experiment section. From now on, we assume the locations of people are given. On the training data, each image is associated with a group activity label, and each person in the image is associated with an action label.

3.1 A Latent Model for Contextual Group Activities

Our goal is to learn a model that jointly captures the group activity, the individual person actions, and the interactions among them. We present a graphical model representing all the information in a unified framework. One important difference between our model and previous work is that in addition to learning the parameters in the graphical model, we also automatically infer the graph structures (see Sec. 3.1.2).

3.1.1 Model Formulation

A graphical representation of the model is shown in Fig. 3.1. We now describe how we model an image I. Let $I_1, I_2, \ldots, I_m$ be the set of persons found in the image I, we extract features $\mathbf{x}$ from the image I in the form of $\mathbf{x} = (x_0, x_1, \ldots, x_m)$, where x_0 is the aggregation of feature descriptors of all the persons in the image (we call it *root feature vector*), and $x_i (i = 1, 2, \ldots, m)$ is the feature vector extracted from the person I_i. Rather than directly using certain raw features (e.g. the HOG descriptor [7]) as the feature vector x_i in our framework, we employ a bag-of-words style representation for the feature descriptor of each person's action. We train a multi-class SVM classifier based on the feature descriptor of each individual and their associated action labels. For example, if we have 40 action categories as in the Collective Activity Dataset, then each feature vector x_i is represented as a 40-dimensional vector, where the k-th entry of this vector is the score of classifying this instance to the k-th class returned by the SVM classifier. The root feature vector x_0 of an image is also represented as a 40-dimensional vector, which is obtained by taking an average over all the feature vectors x_i $(i = 1, 2, ..., m)$ in the same image. We denote the collective actions of all the persons in the image as $\mathbf{h} = (h_1, h_2, \ldots, h_m)$, where $h_i \in \mathcal{H}$ is the action label of the person I_i and $\mathcal{H}$ is the set of all possible action labels. The image I is associated with a group activity label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all possible activity labels.



(a) (b)

Figure 3.1: Graphical illustration of the model in (a). The edges represented by dashed lines indicate the connections are latent. Different types of potentials are denoted by lines with different colors in the example shown in (b).

We assume there are connections between some pairs of action labels (h_j, h_k). Intuitively speaking, this allows the model to capture important correlations between action labels. We use an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent $(h_1, h_2, \ldots, h_m)$, where a vertex $v_i \in \mathcal{V}$ corresponds to the action label h_i, and an edge $(v_j, v_k) \in \mathcal{E}$ corresponds to the interactions between h_j and h_k.

We use $f_w(\mathbf{x}, \mathbf{h}, y; \mathcal{G})$ to denote the compatibility of the image feature $\mathbf{x}$, the collective action labels $\mathbf{h}$, the group activity label y, and the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We assume $f_w(\mathbf{x}, \mathbf{h}, y; \mathcal{G})$ is parameterized by w and is defined as follows:

$$f_w(\mathbf{x}, \mathbf{h}, y; \mathcal{G}) = w^\top \Psi(y, \mathbf{h}, \mathbf{x}; \mathcal{G}) \tag{3.1a}$$

$$= w_0^\top \phi_0(y, x_0) + \sum_{j \in \mathcal{V}} w_1^\top \phi_1(x_j, h_j) + \sum_{j \in \mathcal{V}} w_2^\top \phi_2(y, h_j) + \sum_{j,k \in \mathcal{E}} w_3^\top \phi_3(y, h_j, h_k) \tag{3.1b}$$

The model parameters w are simply the combination of four parts, $w = \{w_1, w_2, w_3, w_4\}$. The details of the potential functions in Eq. 3.1 are described in the following:

Image-Action Potential $w_1^\top \phi_1(x_j, h_j)$: This potential function models the compatibility between the j-th person's action label h_j and its image feature x_j. It is parameterized as:

$$w_1^\top \phi_1(x_j, h_j) = \sum_{b \in \mathcal{H}} w_{1b}^\top \mathbb{1}(h_j = b) \cdot x_j \tag{3.2}$$

where x_j is the feature vector extracted from the j-th person and we use $\mathbb{1}()$ to denote the indicator function. The parameter w_1 is simply the concatenation of w_{1b} for all $b \in \mathcal{H}$.

Action-Activity Potential $w_2^\top \phi_2(y, h_j)$: This potential function models the compatibility between the group activity label y and the j-th person's action label h_j. It is parameterized

as:

$$w_2^\top \phi_2(y, h_j) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{H}} w_{2ab}^\top \mathbb{1}(y = a) \cdot \mathbb{1}(h_j = b) \tag{3.3}$$

Action-Action Potential $w_3^\top \phi_3(y, h_j, h_k)$: This potential function models the compatibility between a pair of individuals' action labels (h_j, h_k) under the group activity label y, where $(j, k) \in \mathcal{E}$ corresponds to an edge in the graph. It is parameterized as:

$$w_3^\top \phi_3(y, h_j, h_k) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{H}} \sum_{c \in \mathcal{H}} w_{3abc}^\top \mathbb{1}(y = a) \cdot \mathbb{1}(h_j = b) \cdot \mathbb{1}(h_k = c) \tag{3.4}$$

Image-Activity Potential $w_0^\top \phi_0(y, x_0)$: This potential function is a root model which measures the compatibility between the activity label y and the root feature vector x_0 of the whole image. It is parameterized as:

$$w_0^\top \phi_0(y, x_0) = \sum_{a \in \mathcal{Y}} w_{0a}^\top \mathbb{1}(y = a) \cdot x_0 \tag{3.5}$$

The parameter w_{0a} can be interpreted as a root filter that measures the compatibility of the class label a and the root feature vector x_0.

Inspired by the latent SVM [12], we define the following function to score an image $\mathbf{x}$ and a group activity label y:

$$F_w(\mathbf{x}, y) = \max_{\mathcal{G}} \max_{\mathbf{h}} f_w(\mathbf{x}, \mathbf{h}, y; \mathcal{G}) = \max_{\mathcal{G}} \max_{\mathbf{h}} w^\top \Psi(\mathbf{x}, \mathbf{h}, y; \mathcal{G}) \tag{3.6}$$

The group activity label of the image $\mathbf{x}$ can be inferred as:

$$y^* = \arg\max_y F_w(\mathbf{x}, y) \tag{3.7}$$

Notice that in Eq. 3.6, we explicitly maximize over the graph $\mathcal{G}$. This is very different from previous work which typically assumes the graph structure is fixed.

3.1.2 Learning and Inference

We now describe how to infer the label given the model parameters, and how to learn the model parameters from a set of training data. If the graph structure $\mathcal{G}$ is known and fixed, we can apply standard learning and inference techniques of latent SVMs. For our application, a good graph structure turns out to be crucial, since it determines which person interacts (i.e. provides action context) with another person. The interaction of individuals turns out to

be important for group activity recognition, and fixing the interaction (i.e. graph structure) using heuristics does not work well. We will demonstrate this experimentally in Sec. 3.3. We instead develop our own inference and learning algorithms that automatically infer the best graph structure from a particular set.

Inference

Given the model parameters w and an example $\mathbf{x}$, we can enumerate all the possible $y \in \mathcal{Y}$ and predict the activity label y^* of $\mathbf{x}$ according to Eq. 3.7. For a graph structure $\mathcal{G}_y$, we need to solve the following inference problem of finding the best $\mathbf{h}_y$:

$$\mathbf{h}_y^* = \arg\max_{\mathbf{h}_y} f_w(\mathbf{x}, \mathbf{h}_y, y; \mathcal{G}_y) = \arg\max_{\mathbf{h}_y} w^\top \Psi(y, \mathbf{h}_y, \mathbf{x}; \mathcal{G}_y) \tag{3.8}$$

We use the subscript y in the notations $\mathbf{h}_y$ and $\mathcal{G}_y$ to emphasize that we are now fixing on a particular activity label y. Note that the graph structure $\mathcal{G}_y$ might be different for different y's.

However, since we do not know the graph structure $\mathcal{G}_y$, we simply treat it as yet another latent variable and maximize over it, i.e.

$$(\mathbf{h}_y^*, \mathcal{G}_y^*) = \arg\max_{\mathbf{h}_y, \mathcal{G}_y} w^\top \Psi(y, \mathbf{h}_y, \mathbf{x}; \mathcal{G}_y) \tag{3.9}$$

The optimization problem in Eq. 3.9 is in general NP-hard since it involves a combinatorial search. We instead use an coordinate ascent style algorithm to approximately solve Eq. 3.9 by iterating the following two steps:

1. Holding the graph structure $\mathcal{G}_y$ fixed, optimize the action labels $\mathbf{h}_y$ for the $\langle \mathbf{x}, y \rangle$ pair:

$$\mathbf{h}_y = \arg\max_{\mathbf{h}'} w^\top \phi(\mathbf{x}, \mathbf{h}', y; \mathcal{G}_y) \tag{3.10}$$

2. Holding $\mathbf{h}_y$ fixed, optimize graph structure $\mathcal{G}_y$ for the $\langle \mathbf{x}, y \rangle$ pair:

$$\mathcal{G}_y = \arg\max_{\mathcal{G}'} w^\top \phi(\mathbf{x}, \mathbf{h}_y, y; \mathcal{G}') \tag{3.11}$$

The problem in Eq. 3.10 is a standard max-inference problem in an undirected graphical model. Here we use loopy belief propagation to approximately solve it. The problem in Eq. 3.11 is still an NP-hard problem since it involves enumerating all the possible graph structures. Even if we can enumerate all the graph structures, we might want to restrict ourselves to a subset of graph structures that will lead to efficient inference (e.g. when

using loopy BP in Eq. 3.10). One obvious choice is to restrict $\mathcal{G}'$ to be a tree-structured graph, since loopy BP is exact and tractable for tree structured models. However, as we will demonstrate in Sec. 3.3, the tree-structured graph built from simple heuristic (e.g. minimum spanning tree) does not work that well. Another choice is to choose graph structures that are "sparse", since sparse graphs tend to have fewer cycles, and loopy BP tends to be efficient in graphs with fewer cycles. In this dissertation, we enforce the graph sparsity by setting a threshold d on the maximum degree of any vertex in the graph. When $\mathbf{h}_y$ is fixed, we can formulate an integer linear program (ILP) to find the optimal graph structure (Eq. 3.11) with the additional constraint that the maximum vertex degree is at most d. Let $z_{jk} = 1$ indicate that the edge (j, k) is included in the graph, and 0 otherwise. The ILP can be written as:

$$\max_{z} \sum_{j \in \mathcal{V}} \sum_{k \in \mathcal{V}} z_{jk} \psi_{jk} \tag{3.12a}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{V}} z_{jk} \leq d, \quad \sum_{k \in \mathcal{V}} z_{jk} \leq d, \quad z_{jk} = z_{kj}, \quad \forall j, k \tag{3.12b}$$

$$z_{jk} \in \{0, 1\}, \quad \forall j, k \tag{3.12c}$$

where we use ψ_{jk} to collectively represent the summation of all the pairwise potential functions in Eq. 3.1 for the pairs of vertices (j, k). Of course, the optimization problem in Eq. 3.12 is still hard due to the integral constraint in Eq. 3.12c. But we can relax Eq. 3.12c with a linear constraint $0 \leq z_{jk} \leq 1$ and solve a linear program (LP) instead. The solution of the LP relaxation might have fractional numbers. To get integral solutions, we simply round them to the closest integers.

Learning

Given a set of N training examples $\langle \mathbf{x}^n, \mathbf{h}^n, y^n \rangle$ $(n = 1, 2, \ldots, N)$, we would like to train the model parameter $\mathbf{w}$ that tends to produce the correct group activity y for a new test image $\mathbf{x}$. Note that the action labels $\mathbf{h}$ are observed on training data, but the graph structure $\mathcal{G}$ (or equivalently the variables $\mathbf{z}$) are unobserved and will be automatically inferred. A natural way of learning the model is to adopt the latent SVM formulation [12, 49] as follows:

$$\min_{w, \xi \geq 0, \mathcal{G}_y} \frac{1}{2}||w||^2 + C \sum_{n=1}^{N} \xi_n \tag{3.13a}$$

$$\text{s.t.} \quad \max_{\mathcal{G}_{y^n}} f_w(\mathbf{x}^n, \mathbf{h}^n, y^n; \mathcal{G}_{y^n}) - \max_{\mathcal{G}_y} \max_{\mathbf{h}_y} f_w(\mathbf{x}^n, \mathbf{h}_y, y; \mathcal{G}_y) \geq \Delta(y, y^n) - \xi_n, \forall n, \forall y \tag{3.13b}$$

where $\Delta(y, y^n)$ is a loss function measuring the cost incurred by predicting y when the ground-truth label is y^n. In standard multi-class classification problems, we typically use the 0-1 loss $\Delta_{0/1}$ defined as:

$$\Delta_{0/1}(y, y^n) = \begin{cases} 1 & \text{if } y \neq y^n \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

The constrained optimization problem in Eq. 3.13 can be equivalently written as an unconstrained problem:

$$\min_{w,\xi} \quad \frac{1}{2}||w||^2 + C\sum_{n=1}^{N} R^n \tag{3.15a}$$

$$\text{where} \quad R^n = \max_{y} \max_{\mathbf{h}_y} \max_{\mathcal{G}_y} (\Delta(y, y^n) + f_w(\mathbf{x}^n, \mathbf{h}_y, y; \mathcal{G}_y)) - \max_{\mathcal{G}_{y^n}} f_w(\mathbf{x}^n, \mathbf{h}^n, y^n; \mathcal{G}_{y^n}) \tag{3.15b}$$

We use the non-convex bundle optimization in [10] to solve Eq. 3.15. In a nutshell, the algorithm iteratively builds an increasingly accurate piecewise quadratic approximation to the objective function. During each iteration, a new linear cutting plane is found via a subgradient of the objective function and added to the piecewise quadratic approximation. Now the key issue is to compute the subgradient of Eq. 3.15 for a particular w. Since $\partial_w(\frac{1}{2}||w||^2 + C\sum_{n=1}^{N} R^n) = w + C\sum_{n=1}^{N} \partial_w R^n$, all we need to do is to figure out how to compute $\partial_w R^n$.

Let $(y^*, \mathbf{h}^*, \mathcal{G}^*)$ be the solution to the following optimization problem:

$$\max_{y} \max_{\mathbf{h}} \max_{\mathcal{G}} \Delta(y, y^n) + f_w(\mathbf{x}^n, \mathbf{h}, y; \mathcal{G}) \tag{3.16}$$

The inference problem in Eq. 3.16 is similar to the inference problem in Eq. 3.9, except for an additional term $\Delta(y, y^n)$. Since the number of possible choices of y is small (e.g.$|\mathcal{Y}| = 5$) in our case), we can enumerate all possible $y \in \mathcal{Y}$ and solve the inference problem in Eq. 3.9 for each fixed y. Similarly, let $\hat{\mathcal{G}}$ be the solution to the following optimization problem:

$$\max_{\mathcal{G}'} f_w(\mathbf{x}^n, \mathbf{h}^n, y^n; \mathcal{G}') \tag{3.17}$$

The problem in Eq. 3.17 can be approximately solved using the LP relaxation of Eq. 3.12. Then we can show that the subgradient $\partial_w R^n$ can be calculated as $\partial_w R^n = \Psi(\mathbf{x}^n, y^*, \mathbf{h}^*; \mathcal{G}^*) - \Psi(\mathbf{x}^n, y^n, \mathbf{h}^n; \hat{\mathcal{G}})$. Using this subgradient, we can optimize Eq. 3.13 using the algorithm in [10].

3.2 A Context Descriptor

As mentioned earlier, our second approach explores the person-person interaction in feature level. Note that this approach still utilizes the latent SVM framework, however, we don't consider any pairwise connections between variables $\mathbf{h}$ in the hidden layer, but focus on attaching contextual information into feature descriptors $\mathbf{x}$.

We develop a novel feature representation called the *action context (AC) descriptor*. Our AC descriptor is centered on a person (the focal person), and describes the action of the focal person and the behavior of other people nearby. For each focal person, we set a spatio-temporal context region around him (see Fig. 3.2(a)), only those people inside the context region (nearby people) are considered. The AC descriptor is computed by concatenating two feature descriptors: one is the action descriptor that captures the focal person's action, and the other one is the context descriptor that captures the behaviour of other people nearby, as illustrated in Fig. 3.2(b,c).

Here the feature descriptor of each person's action is computed by a bag-of-words style representation as introduced in Sec. 3.1.1. We represent the action descriptor of the i-th person as: $F_i = [S_{1i}, S_{2i}, \ldots, S_{Ki}]$, where K is the number of action classes, S_{ki} is the score of classifying the i-th person to the k-th action class returned by the SVM classifier.

Given the i-th person as the focal person, its context descriptor C_i is computed from the action descriptors of people in the context region. Suppose that the context region is further divided into M regions (we call "sub-context regions") in space and time, as illustrated in Fig. 3.2(b), then the context descriptor is represented as a $M \times K$ dimensional vector computed as follows:

$$C_i = \left[\max_{j \in \mathcal{N}_1(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_1(i)} S_{Kj}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{Kj} \right] \tag{3.18}$$

Where $\mathcal{N}_m(i)$ indicates the indices of people in the m-th "sub-context region" of the i-th person.

The AC descriptor for the i-th person is a concatenation of its action descriptor F_i and its context descriptor C_i: $AC_i = [F_i, C_i]$. As there might be numerous people present in a video sequence, we construct AC descriptors centered around each person. In the end, we will gather a collection of AC descriptors, one per person.

Fig. 3.3 shows examples of the action context descriptors on the nursing home dataset. Fig. 3.3(a) and Fig. 3.3(b) are two frames that contain falling. The persons in the red

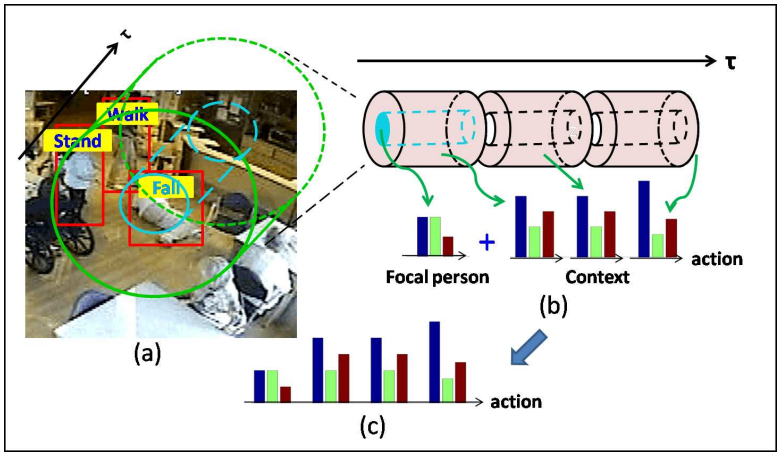


Figure 3.2: Illustration of construction of our action context descriptor. (a) Spatio-temporal context region around focal person, as indicated by the green cylinder. In this example, we regard the fallen person as focal person, and the people standing and walking as context. (b) Spatio-temporal context region around focal person is divided in space and time. The blue region represents the location of the focal person, while the pink regions represent locations of the nearby people. The first 3-bin histogram captures the action of the focal person, which we call the action descriptor. The latter three 3-bin histograms are the context descriptor, and capture the behaviour of other people nearby. (c) The action context descriptor is formed by concatenating the action descriptor and the context descriptor.

bounding boxes are trying to help the fallen residents. Fig 3.3 is a frame that does not contain the falling action. The person in the red bounding box is simply walking across the room. For our application, we would like to distinguish between the high-level activities in Fig. 3.3 (a,b) and Fig. 3.3 (c). However, this is difficult (even for human observers) if we only look at the person in the bounding box, since all three people are walking. But if we look at the context of them, we can easily tell the difference: people in Fig. 3.3 (a,b) are walking to help the fallen residents, while the person in Fig. 3.3 (c) is simply walking. This can be demonstrated by the action context descriptors shown in Fig. 3.3 (d)-(f). Here we use a 20-dimensional action context descriptor and visualize it as a 4×5 matrix so it is easier to compare them visually. We can see that Fig. 3.3 (d) and Fig. 3.3 (e) are similar. Both of them are very different from Fig. 3.3 (f). This demonstrates that the action context descriptor can help us to differentiate people walking to help fallen residents under a *fall* activity from other actions, such as walking under a *nonfall* activity.

The key characteristics of our action context descriptor are in two aspects: 1) instead of simply using features of the neighboring people as context, the action context descriptor employs a bag-of-words style representation which captures the distribution of actions of people nearby. 2) In addition to static context, our descriptor also captures dynamic information, i.e. the temporal evolution of actions extracted from both the focal person and the people nearby.

3.3 Experiments

Most previous work in human action understanding uses standard benchmark datasets to test their algorithms, such as KTH [36] and Weizmann [4] datasets. In the real world, however, the appearance of human activities has tremendous variation due to background clutter, partial occlusion, scale and viewpoint change, etc. The videos in those datasets were recorded in a controlled setting with small camera motion and clean background. The Hollywood human action dataset [23] is more challenging. However, only three action classes: HandShake, HugPerson and Kiss have more than one actor, but these are not contextual – the two actors together perform the one action. (One person does not perform HugPerson by himself.) In this work, we choose to use two challenging datasets to evaluate our proposed method. The first dataset is a benchmark dataset introduced in [6] to study collective human activities. The second dataset consists of surveillance videos collected from

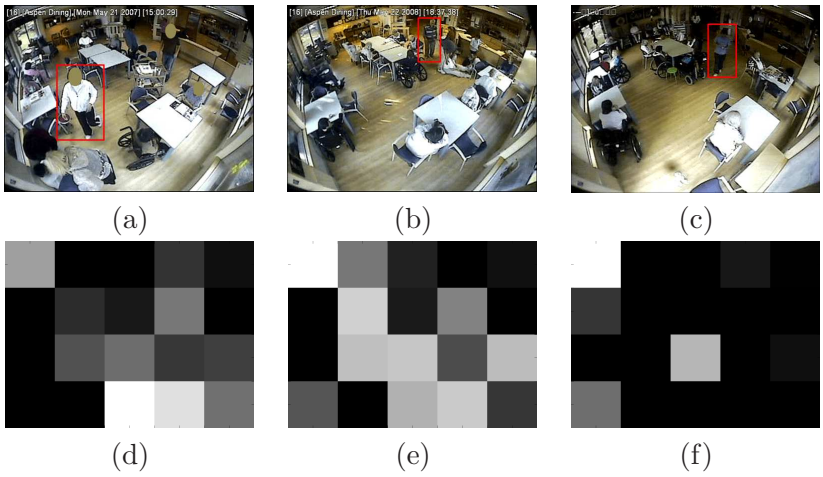
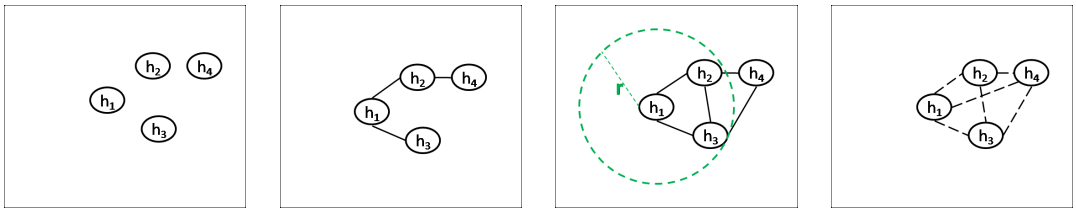


Figure 3.3: Examples of action context descriptors. (a,b) Sample frames containing people falling and other people (shown in red bounding boxes) trying to help the fallen person. (c) A sample frame contain no falling action. The person in the red bounding box is simply walking. (d-f) The action context descriptors for the three persons in bounding boxes. Action context descriptors contain information about the actions of other people nearby.

a nursing home environment by our clinician collaborators.

In our proposed model, we have adopted two approaches to model the person-person interaction: one is in structure level and the other is in feature level, which we call *structure-level approach* and *feature-level approach* respectively from now on. In order to comprehensively evaluate the performance of the proposed model, we compare it with several baseline methods. The first baseline (which we call *root+SVM*) is a SVM model with linear kernel based on the root feature vector x_0, i.e. ignoring the individual actions. The other baselines are within our proposed framework, with various ways of setting the structures of the person-person interaction. The structures we have considered are illustrated in Fig. 3.4(a)-(c), including (a) no pairwise connection; (b) minimum spanning tree; (c) graph obtained by connecting any two vertices within a Euclidean distance ϵ (we call it *ϵ-neighborhood graph*) with $r = 100, 200, 300$. Note that in *structure-level approach* of our proposed model the person-person interactions are latent (shown in Fig. 3.4(d)) and learned automatically. The performance of different structures of person-person interaction are evaluated and compared. We also report the performance of the proposed AC descriptor. In the implementation, we use the AC descriptor to replace the feature vector $x_i (i = 1, 2, \ldots, m)$ in the latent SVM framework. Since the AC descriptor already encodes the person-person interaction, we adopt



(a) (b) (c) (d)

Figure 3.4: Different structures of person-person interaction. Each node here represents a person in a frame. Solid lines represent connections that can be obtained from heuristics. Dashed lines represent latent connections that will be inferred by our algorithm. (a) No connection between any pair of nodes; (b) Nodes are connected by a minimum spanning tree; (c) Any two nodes within a Euclidean distance ϵ are connected (which we call *ϵ-neighborhood graph*); (d) Connections are obtained by structure learning. Note that (d) is the structure of person-person interaction of the proposed *structure-level approach* and our *feature-level approach* employs the structure of (a).

the structure without any pairwise connections, as shown in Fig. 3.4(a). The parameters of the proposed AC descriptor and multi-class SVM are set according to cross-validation in the training set. The regularization constant C in Eq. 3.13 is set emprically in the range of 0.1 to 1.

Person Detectors: As mentioned earlier, how to localize people is task specific. For the Collective Activity Dataset, we apply the pedestrian detector in [12] to find all the candidate regions corresponding to people. The pedestrian detector performs very well on this dataset with only a few false positive detections (see Fig. 3.6(a)). For the Nursing Home dataset, however, pedestrian detectors are not reliable, as shown in Fig. 3.6(b). We instead extract moving regions from the videos as our detected people. First, we perform background subtraction using the OpenCV implementation of the standard Gaussian Mixture Model (GMM) [38] to obtain the foreground regions. Then, we extract all the 8-connected regions of the foreground from each frame, which are considered as moving regions. Moving regions with size less than a threshold *Th* are deemed unreliable and therefore ignored. People's locations in the training set are labeled with bounding boxes, while person detectors are used to localize each person in the test set.

Person Descriptors: We also use different feature descriptors to describe people for the two datasets. HOG descriptor [7] is used for the Collective Activity Dataset. For the nursing home dataset, standard features such as optical flow or HOG [7] are typically not reliable due to low video quality. Instead, we use a feature representation similar to the one

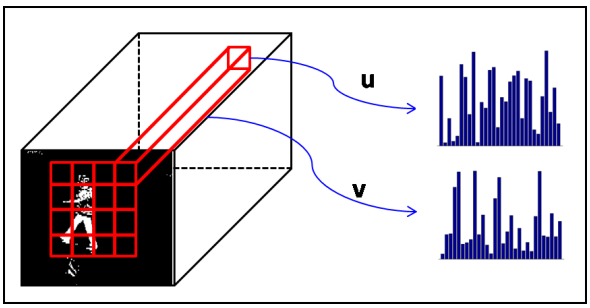


Figure 3.5: Illustration of the local spatio-temporal (LST) feature representation for describing a candidate region. **u** is a vector of percentage of static foreground pixels, **v** is a vector of percentage of moving foreground pixels

introduced in [25], which has been shown to be reliable for low resolution videos. The feature descriptor is computed as follows. We first divide the bounding box a detected person into N blocks. Foreground pixels are detected using standard background subtraction. Each foreground pixel is classified as either static or moving by frame differencing. Each block is represented as a vector composed of two components: $\mathbf{u} = [u_1, \ldots, u_t, \ldots, u_\tau]$ and $\mathbf{v} = [v_1, \ldots, v_t, \ldots, v_\tau]$, where u_t and v_t are the percentage of static and moving foreground pixels at time t respectively. τ is the temporal extent used to represent each moving person. As in [25], we refer to it as local spatio-temporal (LST) descriptor in this dissertation. Fig. 3.5 illustrates the LST descriptor. Note that rather than directly using raw features (e.g. HOG [7] or LST) as the feature vector x_i in our framework, we use the method discussed in Sec. 3.1.1 to reduce feature dimension.

3.3.1 Collective Activity Dataset

This dataset contains 44 video clips acquired using low resolution hand held cameras. In the original dataset, all the persons in every tenth frame of the videos are assigned one of the following five categories: *crossing, waiting, queuing, walking* and *talking*, and one of the following eight pose categories: *right, front-right, front, front-left, left, back-left, back* and *back-right*. Based on the original dataset, we define five activity categories including *crossing, waiting, queuing, walking* and *talking*. We define forty action labels by combining the pose and activity information, i.e. the action labels include *crossing and facing right, crossing and facing front-right,* etc. We assign each frame into one of the five activity categories, by taking the majority of actions of persons (ignoring their pose categories) in

(a) (b)

Figure 3.6: Typical results of running a state-of-the-art pedestrian detector [12] on the two datasets used in the experiments. On the collective activity dataset (a), the detector performs very well. But on the more challenging nursing home dataset (b), the detector is not reliable since the videos are captured by a fish eye camera, so persons in the videos are not in upright positions. In addition, the video quality is very poor.

that frame. We select one fourth of the video clips from each activity category to form the test set, and the rest of the video clips are used for training.

We summarize the comparison on the performance of our approaches and the baselines in Table 3.1. Since the test set is imbalanced, e.g. the number of crossing examples is more than twice that of the queuing or talking examples, we report both overall and mean per-class accuracies. As we can see, for both overall and mean per-class accuracies, our methods (structure-level approach and feature-level approach) achieve the top two performances. The proposed model significantly outperforms *root+SVM*. The confusion matrices of our methods and the baseline *root+SVM* are shown in Fig. 3.7. We can see that by incorporation contextual information (Fig. 3.7(b),(c)), the confusions between crossing, waiting and walking are reduced. This is because the relative facing directions (poses) in a group of people provides useful cues for disambiguate these activities: people always cross street in either the same or the opposite directions; people always wait in the same direction, they rarely wait facing each other; the poses in walking are not as regular as in the previous two activities, people can walk in different directions. These can be further demonstrated by the learned pairwise weights for the five activity classes, as visualized in Fig. 3.8. Besides the poses within the same action class, we can also get which actions tend to occur together in an activity. Generally speaking, the model favors seeing the same actions with different poses together under an activity class, e.g. actions of crossing with different poses are favored under the activity label *crossing*. However, in some cases, several different actions are

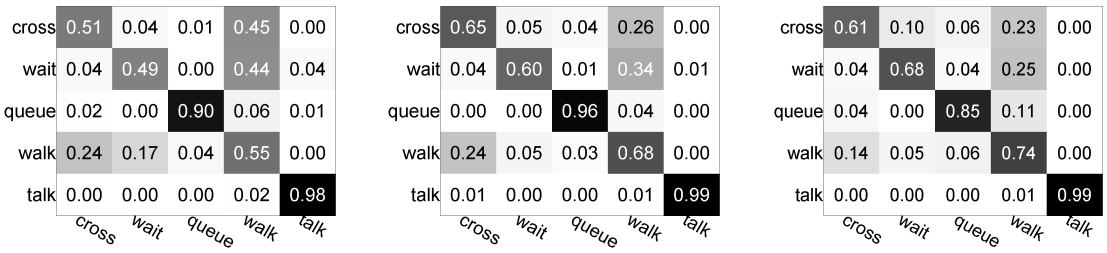


Figure 3.7: Confusion matrices for activity classification on the collective activity dataset: (a) root+SVM. (b) Structure-level approach. (c) Feature-level approach. Rows are ground-truths, and columns are predictions. Each row is normalized to sum to 1.

also favored under the same activity class, e.g. the actions of talking and walking could be together under the activity label *talking*. This is reasonable since when a group of people are talking, some people may pass by.

We visualize the classification results and the learned structure of person-person interaction by *structure-level approach* in Fig. 3.9. Some interesting structures are learnt, like a chain structure which connects people facing the same direction for the *queuing* activity, pairwise connections between people facing the same direction for *waiting* and people facing each other for talking. Note that in the correct classification example of talking, there is a line connects the person in blue and the person in black who are facing the same direction. This is because we made an incorrect prediction of the pose of the person in blue, which are predicted as *front*. Thus, according to our prediction, the connected people (the person in blue and the person in black) are facing each other, thus the learned structure of the talking example is reasonable.

3.3.2 Nursing Home Dataset

Our second dataset consists of videos recorded in a dining room of a nursing home by a low resolution fish eye camera. Typical actions happening in nursing homes include *walking*, *standing*, *sitting*, *bending*, and *falling*. During training, each person is assigned into one of the five action categories. Based on the action categories, we assign each frame into one of the two activity categories: *fall* and *non-fall*. If a frame contains fallen people, then it is labeled as *fall*, otherwise *nonfall*. Our dataset contains one 30-minutes video clip without falls and another thirteen short clips with falls. We divide the dataset into 22 short video

Method	Overall	Mean per-class
root + SVM	70.9	68.6
no connection	75.9	73.7
minimum spanning tree	73.6	70.0
ϵ-neighborhood graph, $\epsilon = 100$	74.3	72.9
ϵ-neighborhood graph, $\epsilon = 200$	70.4	66.2
ϵ-neighborhood graph, $\epsilon = 300$	62.2	62.5
structure-level approach	**79.1**	**77.5**
feature-level approach	**78.5**	**77.5**

Table 3.1: Comparison of activity classification accuracies of different methods on the collective activity dataset. We report both the overall and mean per-class accuracies due to the class imbalance. The first result (root+SVM) is tested in the multi-class SVM framework, while the other results are in the framework of our proposed model but with different structures of person-person interaction. The structures are visualized in Fig. 3.4.

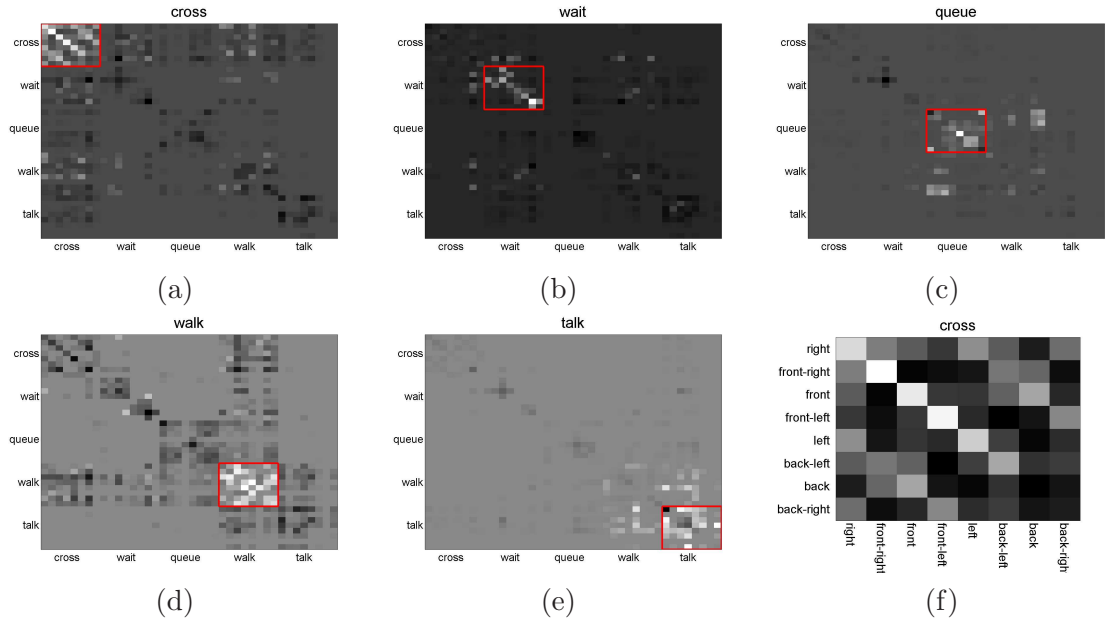


Figure 3.8: Visualization of the weights across pairs of action classes for each of the five activity classes on the collective activity dataset. Light cells indicate large values of weights. Consider the example (a), under the activity label *crossing*, the model favors seeing actions of crossing with different poses together (indicated by the area bounded by the red box). We can also take a closer look at the weights within actions of crossing, as shown in (f). we can see that within the crossing category, the model favors seeing the same pose together, indicated by the light regions along the diagonal. It also favors some opposite poses, e.g. back-right with front-left. These make sense since people always cross street in either the same or the opposite directions.

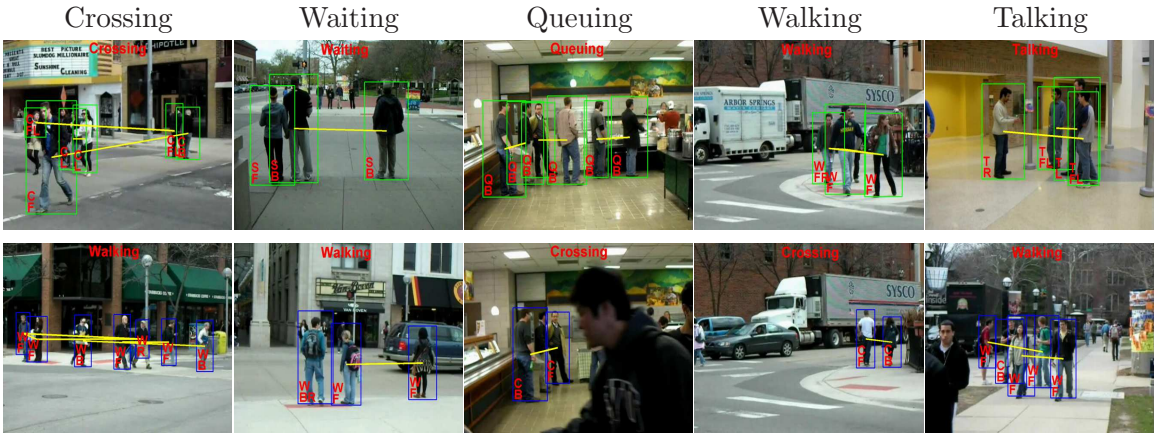


Figure 3.9: (Best viewed in color) Visualization of the classification results and the learned structure of person-person interaction on the collective activity dataset. The top row shows correct classification examples and the bottom row shows incorrect examples. The labels C, S, Q, W, T indicate crossing, waiting, queuing, walking and talking respectively. The labels R, FR, F, FL, L, BL, B, BR indicate right, front-right, front, front-left, left, back-left, back and back-right respectively. The yellow lines represent the learned structure of person-person interaction, from which some important interactions for each activity can be obtained, e.g. a chain structure which connects persons facing the same direction is "important" for the *queuing* activity.

clips, we select 8 clips to form the test set, and the rest of the clips are used for training. In total, there are 2990 annotated frames in the dataset, approximately one third of them have an activity label of fall. We demonstrate the recognition of people falling on this dataset, since this is the most interesting and relevant activity for clinicians.

Our work on activity classification on the nursing home dataset is directly inspired by the application of fall analysis in nursing home surveillance videos. Our clinician partners are studying the causes of falls by elderly residents in order to develop strategies for prevention. This endeavor requires the analysis of a large number of video recordings of falls. Alternatives to vision-based analysis for extracting fall instances from a large amount of footage, such as wearable sensors and self-reporting, are inconvenient and unreliable.

We summarize the comparison on the performance of our approaches and the baselines in Table 3.2. Since the test set is imbalanced, i.e. the number of non-fall examples is more than twice that of the fall examples, we report both overall and mean per-class accuracies as in the first experiment. As we can see, for both overall and mean per-class accuracies, our first approach on exploring the latent structures achieves the best performance. The proposed model significantly outperforms *root+SVM*. Also, our second approach on contextual feature descriptor outperforms the original feature descriptor in the same model (*no connection*). Please note that since we don't consider any pairwise connections in *feature-level approach*, it is not directly comparable to other numbers achieved with different structures of the hidden layer. The learned pairwise weights for the two activity classes are visualized in Fig. 3.11. Several important observation can be obtained such as: under the activity label *nonfall*, the model favors seeing action of sitting together with standing or walking; while under the activity label *fall*, the model favors seeing actions of walking, standing and bending together, which happens when clinicians come to help a fallen resident stand up; the action fall typically does not happen together with fall, since there is at most one fall in each frame in this dataset.

This dissertation mainly deals with multi-class and binary classification problems, where the performance of an algorithm is typically measured by its overall accuracy, and the learning approach used is to directly optimize the overall accuracy by 0-1 loss $\Delta_{0/1}$ defined in Eq. 3.14. However, if the dataset is highly imbalanced, the overall accuracy is not an appropriate metric to measure the performance of an algorithm. A better performance measure is the mean per-class accuracy. In this work, we adopt a new loss function introduced in [44] which properly adjust the loss according to the distribution of the classes on the training

data:

$$\Delta_{bal}(y, y^n) = \begin{cases} \frac{1}{m_p} & \text{if } y \neq y^n \text{ and } y^n = p \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

where m_p is the number of examples with class label p. Suppose that we have N training examples, it is easy to verify that $\sum_{n=1}^{N} \Delta_{bal}(y, y^n)$ directly corresponds to the mean per-class accuracy on the training data. When we use the new loss function $\Delta_{bal}(y, y^n)$, the learning algorithm defined in Eq. 3.13 will try to directly maximize the mean per-class accuracy, instead of the overall accuracy. Our task is to classify the two activity categories: fall and non-fall, and the dataset is biased towards non-fall. If we optimize the overall accuracy, more examples will tend to be classified as the dominant class, i.e. non-fall. This is not compatible with our goal, since the clinicians want to extract a large amount of falling examples from surveillance videos even if some non-fall examples are included. The bias towards non-fall examples would lead to missing of many falls. Consequently, we also report the classification results with Δ_{bal}, which are summarized in Table 3.3. We can reach similar conclusions as from Table 3.2. In particular, the mean per-class accuracies of our models are significantly better. It is also interesting to notice that in most cases, models trained with Δ_{bal} achieve lower overall accuracies than trained with $\Delta_{0/1}$ but higher mean per-class accuracies, which is exactly what we expect.

For the classification task, given a test image $\mathbf{x}$, our models (also the baselines) return $|\mathcal{Y}|$ scores $F_w(\mathbf{x}, y)$, where $y \in |\mathcal{Y}|$. We can use these scores to produce Precision-Recall and ROC curves for the positive class, i.e. fall. The score assigned to $\mathbf{x}$ being the class fall can be defined as $f(\mathbf{x}) = F_w(\mathbf{x}, fall) - F_w(\mathbf{x}, nonfall)$. Fig. 3.10 shows the Precision-Recall and ROC curves of our approaches and the baselines for the fall activity class. The comparison of the corresponding Average Precision (AP) and area under ROC (AUC) measures are summarized in Table 3.4. We can see that for both AP and AUC measures, the proposed *structure-level approach* achieves the best performance, and our *feature-level approach* performs significantly better than the baseline under the same model with the original feature descriptor (*no connection*). The loss function we used here is Δ_{bal} which is more suitable to our task than $\Delta_{0/1}$ as argued in the previous paragraph. Please note that we could incorporate any loss function (e.g. F-measure, area under ROC curve in Pascal VOC challenge [11]) into our learning algorithm defined in Eq. 3.13 depending on different tasks.

We visualize the classification results and the learned structure of person-person interaction by *structure-level approach* in Fig. 3.12. From the correct classification examples

Method	Overall	Mean per-class
root + SVM	52.6	53.9
no connection	58.6	56.0
minimum spanning tree	64.1	60.6
ϵ-neighborhood graph, $\epsilon = 100$	69.6	56.2
ϵ-neighborhood graph, $\epsilon = 200$	69.9	61.4
ϵ-neighborhood graph, $\epsilon = 300$	69.4	62.9
structure-level approach	**71.2**	**65.0**
feature-level approach	**63.4**	**57.7**

Table 3.2: Comparison of activity classification accuracies of different methods with $\Delta_{0/1}$ on the nursing home dataset. We report both the overall and mean per-class accuracies due to the class imbalance. The first result (root+SVM) is tested in the multi-class SVM framework, while the other results are in the framework of our proposed model but with different structures of person-person interaction. The structures are visualized in Fig. 3.4.

(Fig. 3.12(a)-(h)), we can see that in many cases, the fallen person can't be detected because of camera placement as in Fig 3.12(a), occlusion as in Fig 3.12(c),(d), and so on. However, we can still correctly classify the high-level activity by using contextual information. That is to say, given some people standing or bending together, we could predict that there is a *fall* even without seeing the fallen person. In the incorrect classification examples (Fig. 3.12(i)-(p)), many mistakes come from incorrect predictions of actions, e.g. standing people close to the camera are easily predicted as sitting because of the change of aspect ratio, as shown in Fig. 3.12(i),(l), people far from the camera could not be reliably recognized due to low resolution, as shown in Fig. 3.12(j),(o),(p). These observations demonstrate a limitation of our approach: our approach does not show reliable predictions for single person's actions, e.g. Fig. 3.12(i),(j),(m),(p), thus when someone falls by himself and nobody around him, we could not expect accurate predictions.

3.3.3 Discussion

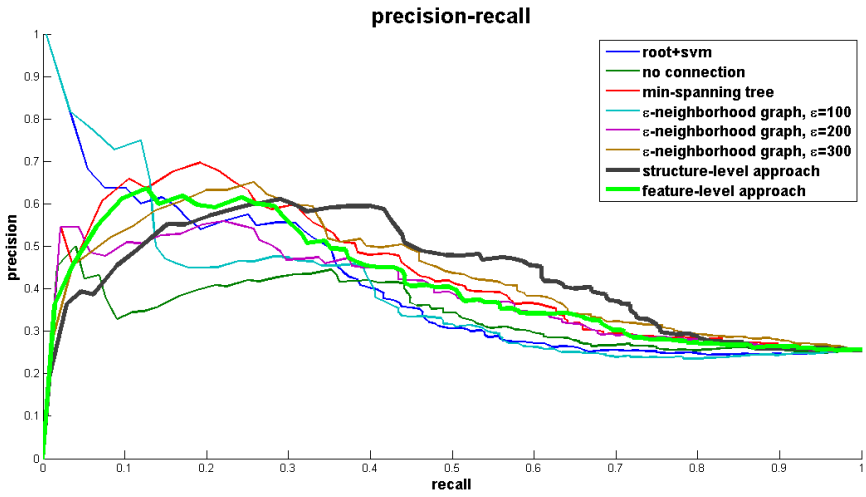
There are several important conclusions we can draw from these experimental results:
Importance of group-person interaction: In the experiments on both of the datasets, no connection between any pair of nodes clearly outperforms *root+SVM*, it is even the best result of the baselines in the experiment on collective activity dataset. It demonstrates the effectiveness of modeling *group-person interaction*, i.e. connection between y and h in our model.

Method	Overall	Mean per-class
root + SVM	48.0	52.4
no connection	54.4	56.1
minimum spanning tree	66.9	62.3
ϵ-neighborhood graph, $\epsilon = 100$	72.7	61.3
ϵ-neighborhood graph, $\epsilon = 200$	67.6	61.1
ϵ-neighborhood graph, $\epsilon = 300$	68.6	64.2
structure-level approach	**71.5**	**67.4**
feature-level approach	**57.3**	**60.3**

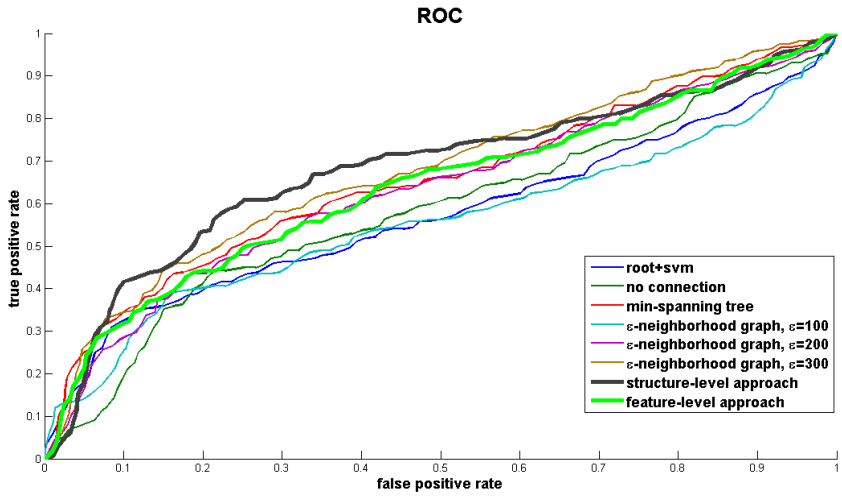
Table 3.3: Comparison of activity classification accuracies of different methods with Δ_{bal} on the nursing home dataset. We report both the overall and mean per-class accuracies due to the class imbalance. The first result (root+SVM) is tested in the multi-class SVM framework, while the other results are in the framework of our proposed model but with different structures of person-person interaction.

Method	AP	AUC
root + SVM	43.3	0.57
no connection	35.8	0.58
minimum spanning tree	45.8	0.65
ϵ-neighborhood graph, $\epsilon = 100$	42.8	0.56
ϵ-neighborhood graph, $\epsilon = 200$	40.2	0.63
ϵ-neighborhood graph, $\epsilon = 300$	45.7	0.67
structure-level approach	**46.6**	**0.68**
feature-level approach	**43.0**	**0.64**

Table 3.4: Comparison of Average Precision (AP) and area under ROC (AUC) measures of different methods on the nursing home dataset. The first result (root+SVM) is tested in the multi-class SVM framework, while the other results are in the framework of our proposed model but with different structures of person-person interaction.

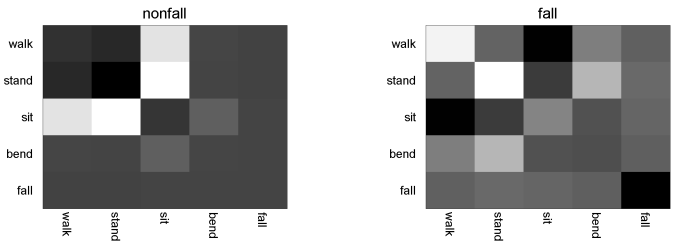


(a)



(b)

Figure 3.10: (Best viewed in color) Comparison of performance for the *fall* activity of different methods in terms of (a) Precision-Recall curves and (b) ROC curves. The comparison of Average Precision (AP) and area under ROC (AUC) measures are shown in Table 3.4.



(a) (b)

Figure 3.11: Visualization of the weights across pairs of action classes for each of the two activity classes on the nursing home dataset. Light cells indicate large values of weights. Consider the example (a), under the activity label *nonfall*, the model favors seeing action of sitting together with standing or walking. These make sense since what usually happen in a non-fall activity are clinicians walking to the sitting residence and standing beside them to offer some help. Typical examples can be referred to Fig. 3.12(e)-(h). Under the activity label *fall*, as shown in (b), the model favors seeing actions of walking, standing and bending together. These usually happen after a residence falls and clinicians come to help the residence stand up. Typical examples are shown in Fig. 3.12(a)-(d). Please note that there is at most one fall in each clip of our dataset, so the action *fall* never happen with *fall*, this is captured by the dark cell in the bottom right corner.

Importance of structure learning of person-person interaction: In Table 3.1, the pre-defined structures such as the minimum spanning tree and the ϵ-*neighborhood graph* do not perform as well as the one without person-person interaction. We believe this is because those pre-defined structures are all based on heuristics and are not properly integrated with the learning algorithm. As a result, they can create interactions that do not help (and sometimes even hurt) the performance. In the experiment on the nursing home dataset, the pre-defined ϵ-*neighborhood graph* achieves better performance than other baselines, as indicated by Table 3.2. We believe this is because of two reasons: first, when a resident falls in a nursing home, most people in the same scene are related to him/her. A common scene is nurses walking to the fallen resident and standing or bending beside the resident to help him/her stand up. Thus a ϵ-neighborhood graph is potentially suitable to this task. Second, the nursing home dataset is collected from real-world surveillance videos, so the video quality is extremely low. Consequently, we could only roughly label five action classes (there are fourty detailed action labels in the collective activity dataset). This would produce fewer outliers that are mistakenly connected by ϵ-neighborhood graph as in the collective activity dataset. If we consider the graph structure as part of our model and

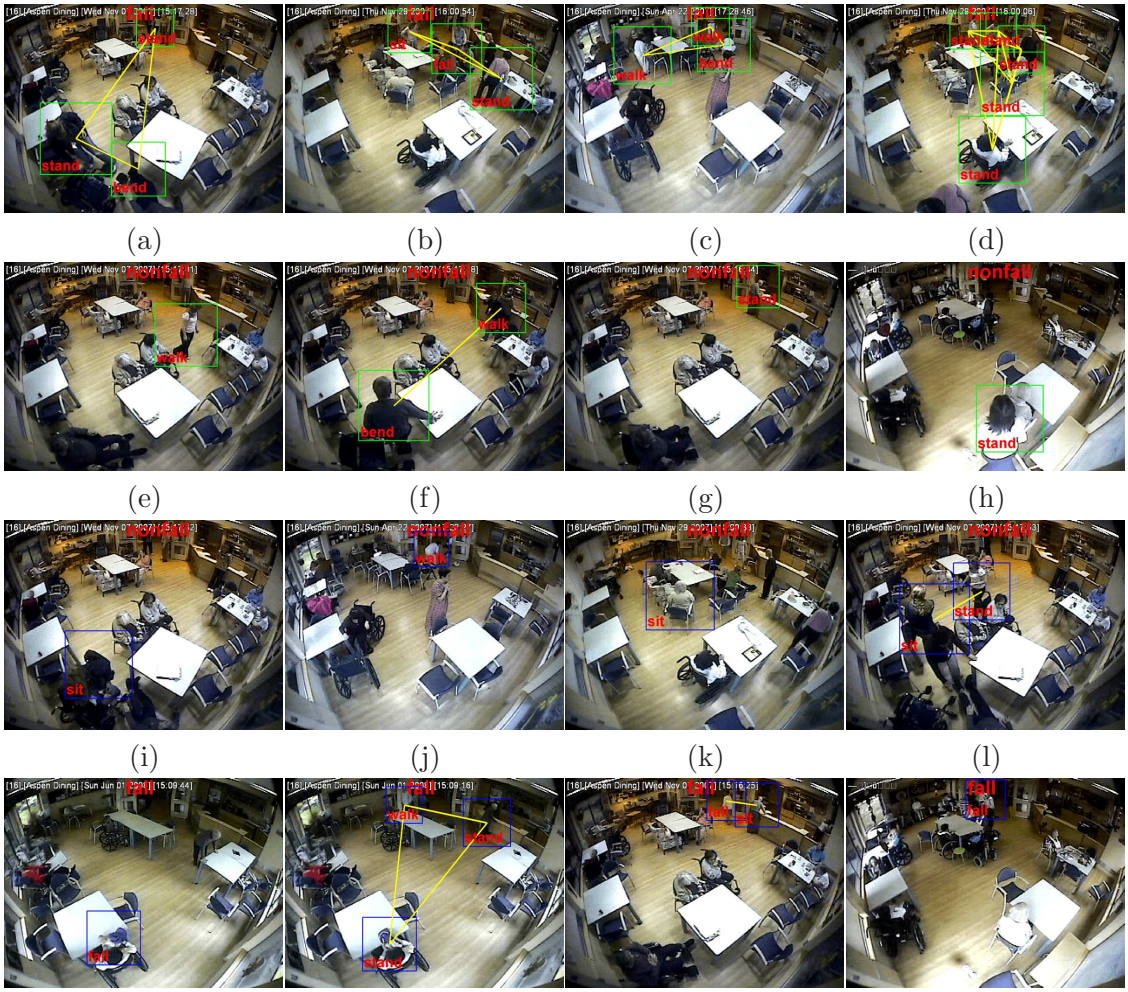


Figure 3.12: (Best viewed in color) Visualization of the classification results and the learned structure of person-person interaction on the nursing home dataset. The first two rows show correct classification examples and the last two rows show incorrect examples. We also show the predicted activity and action labels in each image. The yellow lines represent the learned structure of person-person interaction, from which some important interactions for each activity can be obtained.

directly infer it using our learning algorithm, we can make sure that the obtained structures are those useful for differentiating various activities. Evidence for this is provided by the big jump in terms of the performance by our approach. Even in the experiment on the roughly labeled nursing home dataset, our structure learning algorithm (*structure-level approach*) also achieves better performance than ϵ-*neighborhood graph* which is potentially suitable for this task.

Importance of contextual feature descriptor: In the experiments on both of the datasets, the proposed contextual descriptor significantly outperforms the baseline under the same model with the original feature descriptor (*no connection*).

Chapter 4

Conclusion and Future Work

In this dissertation, we have presented a discriminative model for group activity recognition which jointly captures the group activity, the individual person actions, and the interactions among them. The goal of this dissertation is to demonstrate the effectiveness of contextual information in recognizing group activities. We have exploited two new types of contextual information: *group-person interaction* and *person-person interaction*. In particular, we have proposed two different ways to model the *person-person interaction*, one way is in structure level, we have introduced a structure learning algorithm that automatically infers the optimal structure of person-person interaction in a latent variable framework. The other way is in feature level, we have introduced an action context descriptor that encodes information about action of an individual person in a video, as well as behaviour of other people nearby. In the following, I will briefly highlight the limitations of the current work and several directions for future research.

4.1 Limitations

There is general agreement that context is helpful for visual recognition. However, among all types of contextual information, a great deal of them contribute rather trivial for recognition tasks. Moreover, some of them can be unreliable and ambiguous, thus may not always provide positive effects on visual recognition. How to select the useful context remains unclear. For the two approaches proposed in this dissertation, we will highlight the limitations as follows.

In terms of the structure level approach, our proposed learning algorithm is capable

of learning and optimizing the structure of the hidden layer, which moves a step forward over the most previous latent structured models which assume a predefined structure for the hidden layer, e.g. a tree structure. However, there are still some limitations. For example, as discussed in Sec. 3.1.2, the problem of optimizing graph structures in Eq. 3.11 is NP-hard since it involves enumerating all the possible graph structures. In this work, we enforce the graph sparsity that will lead to more efficient inference (e.g. when using loopy BP). However, this method still could not overcome the intractability because of cycles. Also, several extensions like modeling temporal structures and involving human-object interactions could be explored, detailed discussions are presented in Sec. 4.2.

In terms of the feature level approach, there are two limitations in general. First, the size of contextual regions are set by cross-validation. However, there is no criterion that evaluates how much contextual information should be included that is useful to our task. Second, the useful types of context vary much in different situations: it can be appearance feature of a whole person, a small patch, some attributes, and so on. In this work, we only use the distribution of actions of people nearby as context, in the future, we could also consider other types contextual information and select the most discriminative ones.

4.2 Future Work

The work presented in this dissertation leads to two important directions for future research.
Modeling Complex Structures: The first obvious direction for future work is to solve problems in computer vision that involve complex structures. Latent SVM presents a nice way in modeling structural information of data, for both structured outputs y and structured latent variables h. Consider the problem of human activity recognition. Suppose an activity could be represented by a three-layer model. On the top most layer is the activity y. We could model the structure of y, i.e. the co-occurrence of pairs of activities. Each activity is composed of several actions h, which forms the intermediate layer. The bottom layer is the input video frames x.

The intermediate layer usually involves rich structures. For example, in a baseball activity, typical actions include *pitching*, *hitting*, *missing*, *running*, *catching* and *throwing*. There are temporal dependencies among the actions: e.g. hitting or missing happens after pitching, if batter missed the ball, catching and throwing couldn't happen. Most previous approaches have heavily utilized graphical models with fixed structures. However, the fixed

structures severely limit the adaptability of such models in real-world conditions. The occurrence of actions may not subject to the pre-defined rules, the set of given actions at any given time may vary substantially. In this dissertation, we move a step forward and treat the structure of the intermediate layer as a latent variable and implicitly infer it during learning and inference. Our approach provides a framework of modeling such problems by automatically inferring the optimal structures. As pointed out by the previous section, this framework could be extended in several directions: 1) Temporal dependencies among actions could be modeled. 2) We could introduce the prior structural information into the model that may lead to more efficient inference, e.g. the temporal order of actions as in the baseball game, a chain structure tends to appear in queuing activity, etc. 3) In this work, we only consider the interactions among persons. However, in some cases, the interactions between person and object, person and scene could also provide useful information.

Contextual Feature Descriptors: The sophistication of graphical models leads more challenging learning problems. On the other end of the spectrum, modeling context through feature descriptors is another prominent direction. Our proposed context descriptor encodes information about action of an individual person in a video, as well as behaviour of other people nearby. However, as discussed in Sec. 4.1, there are still several problems that deserve to be addressed in the future. A descriptor that can flexibly encode the discriminative context with a criterion that evaluates the usefulness of different types of context is an interesting direction of future research.

In conclusion, I believe that using context to aid visual recognition is promising in many computer vision applications. It allows us to go beyond single-instance recognition, and explore the mutual dependencies among instances and complex structural information inherent in vast amount of visual data.

Bibliography

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

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

[3] I. Biederman, R. Mezzanotte, and J. Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychol.*, 14(2):143–177, 1982.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision*, 2005.

[5] P. Carbonetto, N. Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.

[6] W. Choi, K. Shahid, and S. Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *9th International Workshop on Visual Surveillance*, 2009.

[7] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[8] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *IEEE International Conference on Computer Vision*, 2009.

[9] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *Workshop on Structured Models in Computer Vision*, 2010.

[10] T. Do and T. Artieres. Large margin training for hidden markov models with partially observed states. In *International Conference on Machine Learning*, 2009.

[11] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[13] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[14] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *European Conference on Computer Vision*, 2008.

[15] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

[16] A. Gupta, P. Srinivasan, J. Shi, and L. Davis. Understanding videos, constructing plots - learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009.

[17] D. Han, L. Bo, and C. Sminchisescu. Selection and context for action recognition. In *IEEE International Conference on Computer Vision*, 2009.

[18] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *European Conference on Computer Vision*, 2008.

[19] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. 80(1):3–15, 2008.

[20] N. Ikizler, R. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *International Conference on Pattern Recognition*, 2008.

[21] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.

[22] S. Khan and M. Shah. Detecting group activities using rigidity of formation. In *ACM Multimedia*, 2005.

[23] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[24] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos"in the wild". In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[25] C. Loy, T. Xiang, and S. Gong. Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In *ICCV*, 2009.

[26] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[27] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889, August 2001.

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

[29] D. Moore and I. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *AAAI*, 2002.

[30] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: A graphicsl model relating features, objects, and scenes. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2004.

[31] A. Chowdhury N. Vaswani and R. Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[32] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852, June 2007.

[33] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *IEEE International Conference on Computer Vision*, 2007.

[34] M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *IEEE International Conference on Computer Vision*, 2009.

[35] M. Ryoo and J. Aggarwal. Stochastic representation and recognition of high-level group activities. *International Journal of Computer Vision*, 2010.

[36] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *17th International Conference on Pattern Recognition*, 2004.

[37] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[38] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[39] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. In *Advances in Neural Information Processing Systems*, volume 17, pages 1401–1408. MIT Press, 2005.

[40] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. 31(3):539–555, 2009.

[41] Y. Wang, H. Jiang, M. Drew, Z. Li, and G. Mori. Unsupervised discovery of action classes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[42] Y. Wang and G. Mori. Human action recognition by semi-latent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision*, 31(10):1762–1774, 2009.

[43] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[44] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision*, 2010.

[45] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69(2), 2006.

[46] T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006.

[47] B. Yao and L. Fei-Fei. Grouplet: a structured image representation for recognizing human and object interactions. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010.

[48] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010.

[49] C. Yu and T. Joachims. Learning structural SVMs with latent variables. In *International Conference on Machine Learning*, 2009.

[50] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings: a two-layer hmm framework. *IEEE Transactions on Multimedia*, 3(8):509–520, 2006.

[51] W. Zheng, S. Gong, and T. Xiang. Quantifying contextual information for object detection. In *IEEE International Conference on Computer Vision*, 2009.

[52] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.