Discriminative Latent Models for Recognizing Contextual Group Activities

Tian Lan, Yang Wang, Weilong Yang, Stephen N. Robinovitch, and Greg Mori, *Member, IEEE*

Abstract—In this paper, we go beyond recognizing the actions of individuals and focus on group activities. This is motivated from the observation that human actions are rarely performed in isolation, the contextual information of what other people in the scene are doing provides a useful cue for understanding high-level activities. We propose a novel framework for recognizing group activities which jointly captures the group activity, the individual person actions, and the interactions among them. Two types of contextual information: *group-person interaction* and *person-person interaction*, are explored in a latent variable framework. In particular, we propose three 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 second approach explores person-person interaction in the 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. The third approach combines the above two. Our experimental results demonstrate the benefit of using contextual information for disambiguating group activities.

1 Introduction

Vision-based human activity recognition is of great scientific and practical importance. Much work in the computer vision literature focuses on single-person action recognition. However, 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. Consider the two persons in Fig. 1(a), can you tell they are doing two different actions? Once the entire contexts of these two images are revealed (Fig. 1(b)) and we observe the interaction of the person with other persons in the group, it is immediately clear that the first person is queuing, while the second person is talking. Another example is from a nursing home surveillance video. The intra-class variation in action categories and relatively poor video quality typical of surveillance footage render this a challenging problem. With this type of video footage many actions are ambiguous, as shown in Fig. 2. 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. In this paper, we argue that actions of individual humans often should not be inferred alone. We instead focus on developing methods for recognizing group activities by modeling the collective behaviors of individuals in the group.

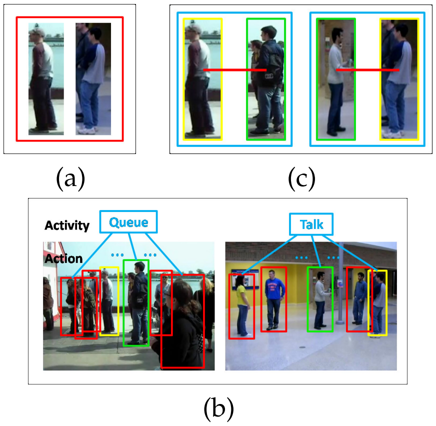


Fig. 1: Role of context in group activities. It is often hard to distinguish actions of each individual person alone (a). However, if we look at the whole scene (b), we can easily recognize the activity of the group and the action of each individual. In this paper, we operationalize this intuition and introduce a model for recognizing group activities jointly considering the group activity, the action of each individual, and the interaction among certain pairs of actions (c).

Before we proceed, we first clarify some terminology used throughout the rest of the paper. 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(b), each frame

- *T. Lan, W. Yang and G. Mori are with the School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.*
  E-mail: tla58,wya16,mori@cs.sfu.ca
- *Y. Wang is with the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA.*
  E-mail: yangwang@uiuc.edu
- *S. Robinovitch is with the School of Kinesiology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.*
  E-mail: stever@sfu.ca

Fig. 2: Sample frames from a nursing home surveillance video. Our goal is to find instances of residents falling down.

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.

Context is critical in recognition for the human visual system [1]. In computer vision, the use of context is also important for solving various recognition problems, especially in situations with poor quality imagery. This is because features are usually not reliable in such circumstances, thus analysis of an individual person or object alone can not yield reliable results. Our proposed approach is based on exploiting two types of contextual information in group activities. 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, 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(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.

In this paper, we develop a latent variable framework for recognizing group activities. Our framework jointly captures the group activity, the individual person actions, and the interactions among them. Person-person interaction is an important cue to understand the group activity, and a straightforward way to model it is to consider the co-occurrence relationships between every pair of persons. However, there are two problems with such an approach. First, this model would induce a dense model with connections between every pair of people, for which exact inference will be intractable. Second, not all people in a scene provide helpful context for disambiguating the action of an individual. Ideally, we would like to consider only those person-person interactions that are important for the group activity. To this end, we propose to use **adaptive structures** that automatically decide on whether the interaction of two persons should be considered. Since this approach is to model the interaction in the structure level, we call it *structure-level approach* in the rest of the paper.

We also propose two other approaches to model the person-person interaction. One approach is to model the person-person interaction in the feature level, which we call *feature-level approach* in the rest of the paper. We propose a context descriptor that encodes information about an individual person in a video, as well as other people nearby. In contrast to the *structure-level approach*, this approach does not consider the high-level inter-label dependencies and thus inference in the model is tractable. The last approach (*combined approach*) integrates the two previous approaches, using the contextual feature descriptor while maintaining the adaptive structures.

We highlight the main contributions of our model. (1) *Group activity:* much 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. (2) *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. (3) *Adaptive structures and context descriptor:* We present three different approaches to model the person-person interaction in structure-level (adaptive structures), feature-level (context descriptor) and both. Portions of this paper appeared previously [2], [3]. Here, we present a unified view of the feature-level [3] and structure-level [2] formulations of group context, a novel combination, and experimental comparisons among them.

The rest of this paper is organized as follows: Section 2 reviews the previous work. Section 3 presents our framework of modeling the group activities. The details of learning and inference of the model are given in Section 4. Section 5 shows our experimental results. Section 6 concludes this paper.

2 Related Work

Using context to aid visual recognition has received much attention recently. 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 [4], objects and objects [5]–[7], objects and so-called "stuff" (amorphous spatial extent, e.g. trees, sky) [8], etc. The work of Jain et al. [7] is close to our work in spirit, which also uses a learnt structure instead of a fully connected model. Unlike their approach which uses a non-parametric model for edge selection, we propose a latent structured model that captures group activity and individual person's actions in a joint framework.

Much previous work in human action recognition focuses on recognizing actions performed by a single person in a video (e.g. [9], [10]). In this setting, there has been work on exploiting context provided by scenes [11] or objects [12]–[14] to help action recognition. In still image action recognition, object-action context [15]–[18] is a popular type of context used for human-object interaction. In this paper, we focus on another type of contextual information – the action-action context, i.e. 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 [19] and AND-OR graphs [20] have been employed. Gupta et al. [20]'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. [21] examine motion and shape features of entire video frames to detect unusual activities. Mehran et al. [22] build a "bag-of-forces" model of the movements of people in a video frame to detect abnormal crowd behavior. The work of Choi et al. [23] is the closest to ours. In that work, person-person context is exploited by a new feature descriptor extracted from a person and its surrounding area.

There is also a line of work on modeling high-level group activities [24]–[32]. Most of the work on group activity focuses on a small range of activities with clear structural information. For example, Vaswani et al. [24] 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. Khan and Shah [25] use rigidity formulation to represent parade activity. They employ a top-down approach which models the entire group as a whole rather than each individual separately. Intille and Bobick [28] use probabilistic techniques for recognizing hand-specified structured activities such as American football plays. Moore and Essa [27] recognize multitasked activities. Cupillard et al. [31] presents an approach for recognizing specific activities such as violence or pickpocketing viewed by several cameras. Chang et al. [32] presents a real-time system to detect aggressive events in prison. Two hierarchical clustering approaches are proposed to group individuals, and events are reasoned at a group level. The main limitation of this line of work is that the models are designed for specific activities with strict rules, e.g. parade, and thus can not be applied to more general activities. Recently, Ryoo and Aggarwal [30] propose a stochastic representation for more general 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 some structure of group activities, and the structures of group activities are learnt automatically.

Our model is directly inspired by some recent work on learning discriminative models that allow the use of latent variables [16], [33]–[36], particularly when the latent variables have complex structures. These models have been successfully applied in many applications in computer vision, e.g. object detection [37], [38], action recognition [35], [39], human-object interaction [16], objects and attributes [40], human poses and actions [41], image region and tag correspondence [42], etc. So far only applications where the structures of latent variables are fixed have been considered, e.g. a tree-structure in [35], [37]. However in our applications, the structures of latent variables are not fixed and have to be inferred automatically.

3 Modeling Contextual Group Activities

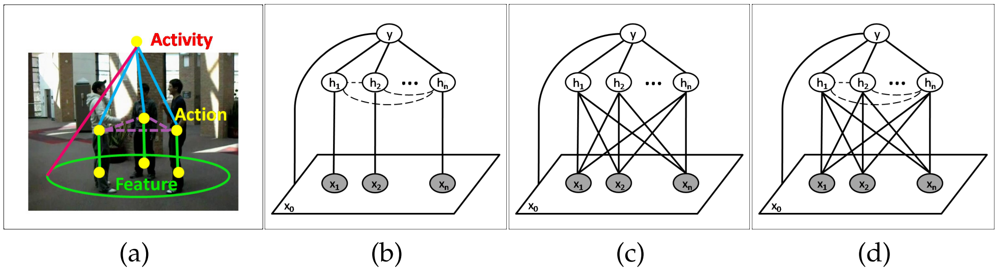
The main objective of our work is to evaluate the benefit of contextual information in group activity recognition. We propose a unified framework that encodes two new types of contextual information, *group-person interaction* and *person-person interaction*. Group-person interaction represents the co-occurrence between the activity of a group and the actions of all the individuals. For example, given a group of people is talking, the action of an individual in the scene is more likely to be talking (whether facing right or left) instead of crossing street. Person-person interaction indicates that the action of an individual can benefit from knowing the actions of other people in the same scene.

We propose three ways to model the person-person interaction: one way is to explore the structures of all pairs of actions, i.e. the structure-level approach; another way is to propose a feature descriptor that captures both the action of an individual person and the behaviour of other people nearby, i.e. the feature-level approach; the third way is to combine the two above mentioned approaches.

3.1 Model Formulation

We assume an image has been pre-processed so the persons in the image have been found. Detecting people in the video frames is task-specific (e.g. [37] or background subtraction), the details are described in the experiments 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.

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 *global feature vector*), and x_i $(i = 1, 2, \ldots, m)$



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

Fig. 3: (a) Illustration of our model (b) on an image of people talking. The edges represented by dashed lines indicate the connections are latent. Different types of potentials are denoted by lines with different colors. (b), (c), (d) are graphical illustrations of our models for the structure-level approach, feature-level approach and combined approach respectively.

is the feature vector extracted from the person I_i. 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.

Fig. 3 shows graphical representations of the three models. We can see that they are in a unified latent structured framework, differing in the way to encode contextual information.

In the *structure-level approach* (Fig. 3(b)), we assume there are connections between some pairs of action labels (h_j, h_k). 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. Different from most of the previous work in latent structured models that 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, this adaptive structure approach will automatically decide on whether the interaction of two persons should be considered, i.e. only the important interactions between people for the recognition task are considered.

For the *feature-level approach* (Fig. 3(c)), we use a similar model, the only difference is that there are no connections between variables $\mathbf{h}$ in the hidden layer – context is attained via features describing the actions of neighboring people. Intuitively, this model encodes correlations among action classes and the contextual feature descriptors that are constructed by the original feature descriptors $\mathbf{x}$. One benefit of including feature-level context is that it does not complicate inference.

In the *combined approach* (Fig. 3(d)), we use the contextual descriptor from the feature-level approach, while maintaining the inter-label dependencies from the structure-level approach.

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 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Note that for the *feature-level approach* and *combined approach*, the feature vector for each person is actually a function of the original feature vectors $\mathbf{x}$, which will be introduced in the next section. Here we use the notation $\mathbf{x}$ for simplicity.

We assume $f_w(\mathbf{x}, \mathbf{h}, y; \mathcal{G})$ is parameterized by w and is defined as follows:

$$\begin{aligned} & f_w(\mathbf{x}, \mathbf{h}, y; \mathcal{G}) = w^\top \Psi(y, \mathbf{h}, \mathbf{x}; \mathcal{G}) \\ & = w_0^\top \phi_0(y, x_0) + \sum_{j \in \mathcal{V}} w_1^\top \phi_1(x_j, h_j) \\ & + \sum_{j \in \mathcal{V}} w_2^\top \phi_2(y, h_j) + \sum_{j,k \in \mathcal{E}} w_3^\top \phi_3(y, h_j, h_k) \end{aligned} \quad (1)$$

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. 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:

$$w_1^\top \phi_1(x_j, h_j) = \sum_{b \in \mathcal{H}} w_{1b}^\top \mathbb{1}(h_j = b) \cdot x_j \quad (2)$$

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:

$$w_2^\top \phi_2(y, h_j) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{H}} w_{2ab} \cdot \mathbb{1}(y = a) \cdot \mathbb{1}(h_j = b) \quad (3)$$

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) and the group activity label y, where $(j, k) \in \mathcal{E}$ corresponds to an edge in the graph. Note that only the models in Fig. 3(b),(d) have this pairwise term. It is parameterized as:

$$\begin{aligned} & w_3^\top \phi_3(y, h_j, h_k) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{H}} \sum_{c \in \mathcal{H}} w_{3abc} \cdot \mathbb{1}(y = a) \\ & \cdot \mathbb{1}(h_j = b) \cdot \mathbb{1}(h_k = c) \end{aligned} \quad (4)$$

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

$$w_0^\top \phi_0(y, x_0) = \sum_{a \in \mathcal{Y}} w_{0a}^\top \mathbb{1}(y = a) \cdot x_0 \tag{5}$$

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

As stated previously, for the feature-level approach and combined approach, we introduce a contextual feature descriptor to replace the original feature vectors $\mathbf{x}$ in Eq. 1. Now we will provide the details on the contextual descriptor in the following section.

3.2 A Contextual Feature Descriptor

In this section, we describe how to encode contextual information into feature descriptors $\mathbf{x}$. This is used by the *feature-level approach* and *combined approach*. Our approach enables analyzing human actions by looking at contextual information extracted from the behaviour of nearby people. A representative example is shown in Fig. 2. With the surveillance video footage, many actions are ambiguous, e.g. falling down and sitting down. A helpful cue to disambiguate these two actions is the context of what other people in the video are doing. 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.

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. 4(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. 4(b,c).

Here we employ a bag-of-words style representation for the action descriptor of each person, which is built in a two-stage approach. First, we train a multi-class SVM classifier based on the person descriptors (e.g. HOG [43]) and their associated action labels. We then represent each person as a K-dimensional vector (i.e. the action descriptor), where K is the number of action classes. The action descriptor of the i-th person is: $F_i = [S_{1i}, S_{2i}, \ldots, S_{Ki}]$, where 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. 4(b), then the context descriptor C_i is represented as a $M \times K$ dimensional vector computed as follows:

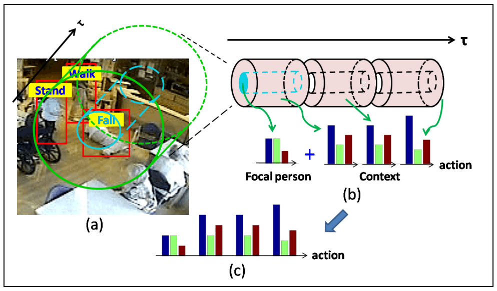


Fig. 4: Illustration of construction of our action context descriptor. (a) Spatio-temporal context region around focal person, as indicated by the green cylinder. In this example, we regard the fallen person as focal person, and the people standing and walking as context. (b) Spatio-temporal context 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.

$$\begin{aligned} C_i = \Big[&\max_{j \in \mathcal{N}_1(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_1(i)} S_{Kj}, \ldots, \\ &\max_{j \in \mathcal{N}_M(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{Kj} \Big] \end{aligned} \tag{6}$$

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. For the *feature-level approach* and *combined approach*, we replace the original feature descriptor x_j in Eq. 2 with the AC descriptor AC_i.

Fig. 5 shows examples of the action context descriptors on the nursing home dataset. Fig. 5(a) and Fig. 5(b) are two frames that contain falling. The persons in the red bounding boxes are trying to help the fallen residents. Fig. 5 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. 5 (a,b) and Fig. 5 (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. 5 (a,b) are walking to help the fallen residents, while the person in Fig. 5 (c) is simply walking. This can be demonstrated by the action context descriptors shown in Fig. 5 (d)-(f). Here we use a 20-

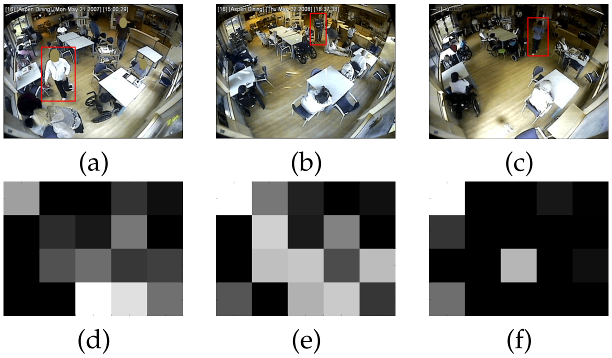


Fig. 5: Examples of action context descriptors. (a,b) Sample frames containing people falling and other people (shown in red bounding boxes) trying to help the fallen person. (c) A sample frame contain no falling action. The person in the red bounding box is simply walking. (d-f) The action context descriptors for the three persons in bounding boxes. Action context descriptors contain information about the actions of other people nearby.

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. 5 (d) and Fig. 5 (e) are similar. Both of them are very different from Fig. 5 (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 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.

4 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. 5. We instead develop our own inference and learning algorithms that automatically infer the best graph structure from a particular set.

4.1 Inference

Given the model parameters w, the inference problem is to find the best group activity label y^* for a new image $\mathbf{x}$. Inspired by the latent SVM [37], we define the following function to score an image $\mathbf{x}$ and a group activity label y:

$$\begin{aligned} F_w(\mathbf{x}, y) &= \max_{\mathcal{G}_y} \max_{\mathbf{h}_y} f_w(\mathbf{x}, \mathbf{h}_y, y; \mathcal{G}_y) \\ &= \max_{\mathcal{G}_y} \max_{\mathbf{h}_y} w^\top \Psi(\mathbf{x}, \mathbf{h}_y, y; \mathcal{G}_y) \end{aligned} \tag{7}$$

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. The group activity label of the image $\mathbf{x}$ can be inferred as: $y^* = \arg\max_y F_w(\mathbf{x}, y)$. Since we can enumerate all the possible $y \in \mathcal{Y}$ and predict the activity label y^* of $\mathbf{x}$, the main difficulty of solving the inference problem is the maximization over $\mathcal{G}_y$ and $\mathbf{h}_y$ according to Eq. 7. Note that in Eq. 7, we explicitly maximize over the graph $\mathcal{G}$. This is very different from previous work which typically assumes the graph structure is fixed.

The optimization problem in Eq. 7 is in general NP-hard since it involves a combinatorial search. We instead use a coordinate ascent style algorithm to approximately solve Eq. 7 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:

$$\mathbf{h}_y = \arg\max_{\mathbf{h}'} w^\top \phi(\mathbf{x}, \mathbf{h}', y; \mathcal{G}_y) \tag{8}$$

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

$$\mathcal{G}_y = \arg\max_{\mathcal{G}'} w^\top \phi(\mathbf{x}, \mathbf{h}_y, y; \mathcal{G}') \tag{9}$$

The problem in Eq. 8 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. 9 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. 8). 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. 5, the tree-structured graph built from a simple heuristic (e.g. minimum spanning tree) does not work 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. A simple way is to include edges if a positive weight is associated with that interaction, and exclude edges with a negative weight. This will create a sparse graph if most of the pairwise interaction weights are not positive. However, sparsity is not guarranteed, since people may interact strongly with each other in some activities. In this paper, 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. 9) 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:

$$\max_{z} \sum_{j\in\mathcal{V}} \sum_{k\in\mathcal{V}} z_{jk}\psi_{jk} \tag{10a}$$

$$\text{s.t.} \sum_{j\in\mathcal{V}} z_{jk} \le d, \sum_{k\in\mathcal{V}} z_{jk} \le d, \ z_{jk} = z_{kj}, \forall j, k \tag{10b}$$

$$z_{jk} \in \{0, 1\}, \quad \forall j, k \tag{10c}$$

where we use ψ_{jk} to collectively represent the summation of all the pairwise potential functions in Eq. 1 for the pairs of vertices (j, k). Of course, the optimization problem in Eq. 10 is still hard due to the integral constraint in Eq. 10c. But we can relax Eq. 10c with a linear constraint $0 \le z_{jk} \le 1$ 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.

4.2 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 [36], [37] as follows:

$$\begin{aligned} &\min_{w,\xi\ge 0,\mathcal{G}_y} \ \frac{1}{2}||w||^2 + C\sum_{n=1}^{N} \xi_n \\ &\text{s.t.} \max_{\mathcal{G}_{y^n}} f_w(\mathbf{x}^n, \mathbf{h}^n, y^n; \mathcal{G}_{y^n}) - \max_{\mathcal{G}_y} \max_{\mathbf{h}_y} f_w(\mathbf{x}^n, \mathbf{h}_y, y; \mathcal{G}_y) \\ &\ge \Delta(y, y^n) - \xi_n, \forall n, \forall y \end{aligned} \tag{11}$$

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:

$$\Delta_{0/1}(y, y^n) = \begin{cases} 1 & \text{if } y \ne y^n \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

The constrained optimization problem in Eq. 11 can be equivalently written as an unconstrained problem:

$$\begin{aligned} &\min_{w,\xi} \quad \frac{1}{2}||w||^2 + C\sum_{n=1}^{N}(\mathcal{L}^n - \mathcal{R}^n) \\ &\text{where } \mathcal{L}^n = \max_{y} \max_{\mathbf{h}_y} \max_{\mathcal{G}_y} (\Delta(y, y^n) + f_w(\mathbf{x}^n, \mathbf{h}_y, y; \mathcal{G}_y)), \\ &\quad \mathcal{R}^n = \max_{\mathcal{G}_{y^n}} f_w(\mathbf{x}^n, \mathbf{h}^n, y^n; \mathcal{G}_{y^n}) \end{aligned} \tag{13}$$

We use the non-convex bundle optimization in [44] to solve Eq. 13. 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 two subgradients $\partial_w \mathcal{L}^n$ and $\partial_w \mathcal{R}^n$ for a particular w, which we describe in detail below.

First we describe how to compute $\partial_w \mathcal{L}^n$. Let $(y^*, \mathbf{h}^*, \mathcal{G}^*)$ be the solution to the following optimization problem:

$$\max_{y} \max_{\mathbf{h}} \max_{\mathcal{G}} \Delta(y, y^n) + f_w(\mathbf{x}^n, \mathbf{h}, y; \mathcal{G}) \tag{14}$$

Then it is easy to show that the subgradient $\partial_w \mathcal{L}^n$ can be calculated as $\partial_w \mathcal{L}^n = \Psi(\mathbf{x}^n, y^*, \mathbf{h}^*; \mathcal{G}^*)$. The inference problem in Eq. 14 is similar to the inference problem in Eq. 7, 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. 7 for each fixed y.

Now we describe how to compute $\partial_w \mathcal{R}^n$, let $\hat{\mathcal{G}}$ be the solution to the following optimization problem:

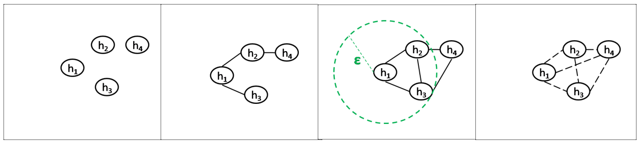
$$\max_{\mathcal{G}'} f_w(\mathbf{x}^n, \mathbf{h}^n, y^n; \mathcal{G}') \tag{15}$$

Then we can show that the subgradient $\partial_w \mathcal{R}^n$ can be calculated as $\partial_w \mathcal{R}^n = \Psi(\mathbf{x}^n, y^n, \mathbf{h}^n; \hat{\mathcal{G}})$. The problem in Eq. 15 can be approximately solved using the LP relaxation of Eq. 10. Using the two subgradients $\partial_w \mathcal{L}^n$ and $\partial_w \mathcal{R}^n$, we can optimize Eq. 11 using the algorithm in [44].

5 Experiments

Most previous work in human action understanding uses standard benchmark datasets to test their algorithms, such as the KTH [10] and Weizmann [9] 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 [45] 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 for collective human activities [23]. The second dataset consists of surveillance videos collected from a nursing home environment by our clinician collaborators.

In order to comprehensively evaluate the performance of the proposed models, we compare them with several baseline methods. The first baseline (which we call *global bag-of-words*) is a SVM model



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

Fig. 6: 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 using adaptive structures. 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).

with linear kernel based on the global feature vector x_0 with a bag-of-words style representation. 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. 6(a)-(c), including (a) no pairwise connection; (b) minimum spanning tree; (c) graph obtained by connecting any two vertices within a Euclidean distance ε (*ε-neighborhood graph*) with $\varepsilon = 100, 200, 300$ and $\inf$ (complete graph). Note that in *structure-level approach* of our proposed model the person-person interactions are latent (shown in Fig. 6(d)) and learned automatically. The performance of different structures of person-person interaction are evaluated and compared. We also report the performance of the feature-level approach and combined approach. 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. 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. 11 is set empirically in the range of 0.1 to 10.

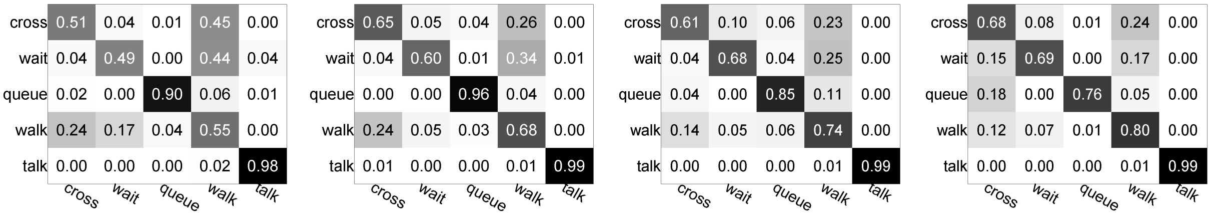
Person Detectors: As mentioned earlier, how to localize people is task specific. For the Collective Activity Dataset, we apply the pedestrian detector in [37]. For the Nursing Home dataset, however, pedestrian detectors are not reliable. We instead extract moving regions from the videos as our detected people. First, we perform background subtraction using the OpenCV implementation of the standard Gaussian Mixture Model (GMM) [46] 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. Person locations in the training set are manually labeled with bounding boxes, while person detectors are used to automatically localize each person in the test set.

Person Descriptors: We also use different feature descriptors to describe people for the two datasets. HOG descriptor [43] is used for the Collective Activity Dataset. For the nursing home dataset, standard features such as optical flow or HOG [43] are typically not reliable due to low video quality. Instead, we use a feature representation similar to the one introduced in [47], which has been shown to be reliable for low resolution videos. The feature descriptor is computed as follows. We first divide the bounding box of 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 [47], we refer to it as local spatio-temporal (LST) descriptor in this paper. Note that rather than directly using raw features (e.g. HOG [43] or LST) as the feature vector x_i in our framework, we use a bag-of-words style representation discussed in Sec. 3.2 to reduce feature dimension.

5.1 Collective Activity Dataset

This dataset contains 44 video clips acquired using low resolution hand-held cameras. In the original dataset, all the people 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 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 of our approaches and the baselines in Table 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, feature-level approach and combined approach) achieve the top three performances. The proposed models significantly outperform *global bag-of-words*. The confusion matrices of our methods and the baseline *global bag-of-words* are shown in Fig. 7. We



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

Fig. 7: Confusion matrices for activity classification on the collective activity dataset: (a) Global bag-of-words. (b) Structure-level approach. (c) Feature-level approach. (d) Combined approach. Rows are ground-truths, and columns are predictions. Each row is normalized to sum to 1.

can see that by incorporation contextual information (Fig. 7(b),(c),(d)), 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 the street in either the same or 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. 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 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. 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 that 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.

5.2 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 include *walking*, *standing*, *sitting*, *bending*, and *falling*. During training, each person is assigned into one of these 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-minute video clip without falls and another thirteen short clips with falls. The frame rate of the video clips is 3 fps. We divide the dataset into 22 short video 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.

Method	Overall	Mean per-class
global bag-of-words	70.9	68.6
no connection	75.9	73.7
minimum spanning tree	73.6	70.0
ε-neighborhood graph, ε = 100	74.3	72.9
ε-neighborhood graph, ε = 200	70.4	66.2
ε-neighborhood graph, ε = 300	62.2	62.5
complete graph	62.6	58.7
structure-level approach	**79.1**	**77.5**
feature-level approach	**78.5**	**77.5**
combined approach	**79.7**	**78.4**

TABLE 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 (global bag-of-words) 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. 6.

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 of our approaches and the baselines in Table 2. Again, we report both

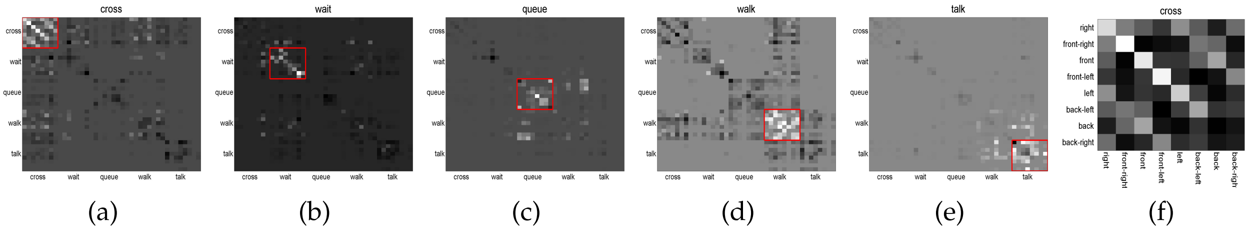


Fig. 8: Visualization of the weights across pairs of action classes for each of the five activity classes on the collective activity dataset. Light cells indicate large values of weights. Consider the example (a), under the activity label *crossing*, the model favors seeing actions of crossing with different 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.

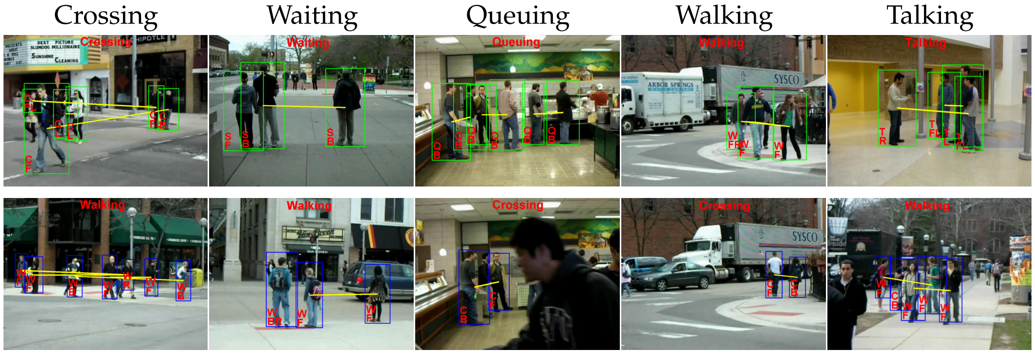


Fig. 9: (Best viewed in color) Visualization of the classification results and the learned structure of person-person interaction on the collective activity dataset. The top row shows correct classification examples and the bottom row shows incorrect examples. The labels C, S, Q, W, T indicate crossing, 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.

overall and mean per-class accuracies since the classes are imbalanced. For both overall and mean per-class accuracies, the proposed models significantly outperform *global bag-of-words*. Also, our second approach using a contextual feature descriptor outperforms the original feature descriptor in the same model (*no connection*). 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. And we can see the clear performance increase by including adaptive structures. The learned pairwise weights for the two activity classes are visualized in Fig. 10. Several important observations 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 staff bend 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 paper 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. 12. 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 the loss function introduced in [40] which properly adjust the loss according to the distribution of the classes on the training data:

$$\Delta_{bal}(y, y^n) = \begin{cases} \frac{1}{m_p} & \text{if } y \neq y^n \text{ and } y^n = p \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

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. 11 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

Method	Overall	Mean per-class
global bag-of-words	52.6	53.9
no connection	58.6	56.0
minimum spanning tree	64.1	60.6
ε-neighborhood graph, ε = 100	69.6	56.2
ε-neighborhood graph, ε = 200	69.9	61.4
ε-neighborhood graph, ε = 300	69.4	62.9
complete graph	70.0	63.1
structure-level approach	**71.2**	**65.0**
feature-level approach	**63.4**	**57.7**
combined approach	**74.3**	**62.3**

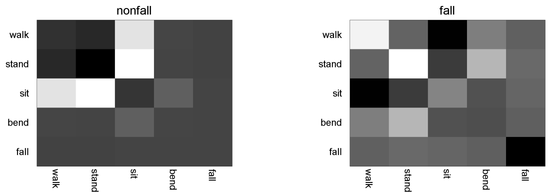
TABLE 2: Comparison of activity classification accuracies of different methods with $\Delta_{0/1}$ on the nursing home dataset. We report both the overall and mean per-class accuracies due to the class imbalance. The first result (global bag-of-words) is tested in the multi-class SVM framework, while 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. 6.

Method	Overall	Mean per-class
global bag-of-words	48.0	52.4
no connection	54.4	56.1
minimum spanning tree	66.9	62.3
ε-neighborhood graph, ε = 100	72.7	61.3
ε-neighborhood graph, ε = 200	67.6	61.1
ε-neighborhood graph, ε = 300	68.6	64.2
complete graph	70.6	62.2
structure-level approach	**71.5**	**67.4**
feature-level approach	**57.3**	**60.3**
combined approach	**69.2**	**63.9**

TABLE 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 (global bag-of-words) 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.

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. We can reach similar conclusions as from Table 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. 11 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 4. We can see that for both



(a) (b)

Fig. 10: 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. 12(c),(d). 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 staff come to help the residence stand up. Typical examples are shown in Fig. 12(a),(b). 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.

AP and AUC measures, the proposed *combined approach* and *structure-level approach* achieve the top two performances, 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. Note that we could incorporate any loss function (e.g. F-measure, area under ROC curve in Pascal VOC challenge [48]) into our learning algorithm defined in Eq. 11 depending on different tasks.

We visualize the classification results and the learned structure of person-person interaction by *structure-level approach* in Fig. 12. From the correct classification examples (Fig. 12(a)-(d)), we can see that in many cases, the fallen person can't be detected because of camera placement, occlusion, and so on (see Fig. 12(a)). 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. 12(e)-(h)), 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 (Fig. 12(f)), people far from the camera could not be reliably recognized due to low resolution (Fig. 12(e),(h)). These observations demonstrate a limitation of our approach: our approach does not show reliable predictions for single person's actions, thus when someone falls by himself with nobody around him, we do not necessarily expect accurate predictions.

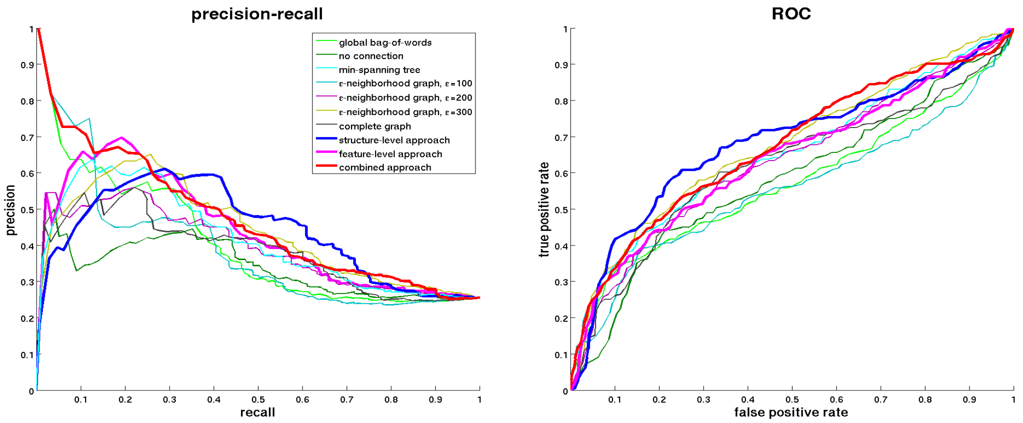
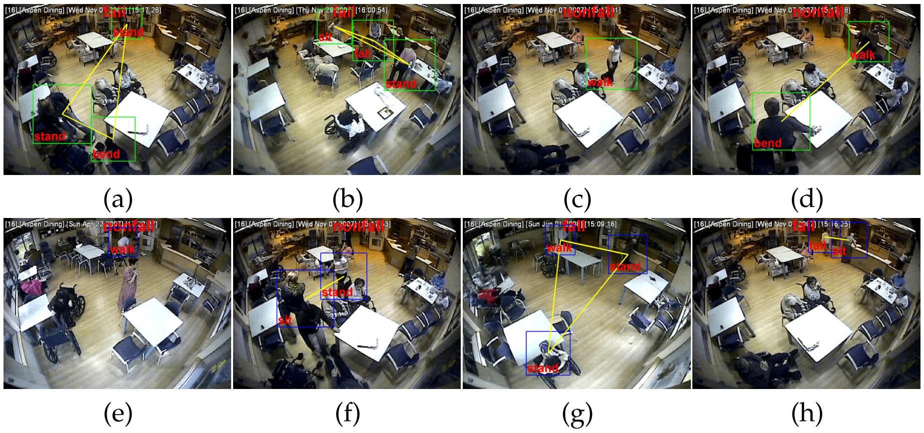


Fig. 11: (Best viewed in color) Comparison of performance for the *fall* activity of different methods in terms of Precision-Recall curves (left) and ROC curves (right). The comparison of Average Precision (AP) and area under ROC (AUC) measures are shown in Table 4.



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

(e) (f) (g) (h)

Fig. 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 row shows correct classification examples and the last row shows 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.

Method	AP	AUC
global bag-of-words	43.3	0.57
no connection	35.8	0.58
minimum spanning tree	45.8	0.65
ε-neighborhood graph, ε = 100	42.8	0.56
ε-neighborhood graph, ε = 200	40.2	0.63
ε-neighborhood graph, ε = 300	45.7	0.67
complete graph	40.1	0.62
structure-level approach	**46.6**	**0.68**
feature-level approach	**43.0**	**0.64**
combined approach	**48.8**	**0.67**

TABLE 4: Comparison of Average Precision (AP) and area under ROC (AUC) measures of different methods on the nursing home dataset. The first result (global bag-of-words) 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.

5.3 Discussion

There are several important conclusions we can draw from these experimental results:

Importance of context in group activity recognition: In the experiments on both of the datasets, our models and all of the baselines with structures clearly outperforms *global bag-of-words*. It demonstrates the effectiveness of modeling *group-person interaction* and *person-person interaction*.

Comparison of adaptive structures and fixed structures: In Table 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. The poor performance of the approximate algorithm in the dense graph is another concern.

In the experiment on the nursing home dataset, the pre-defined *ε-neighborhood graph* achieves better performance than other baselines, as indicated by Table 2. We believe this is for three reasons: first, when a resident falls in a nursing home, most people in the same scene are related to him/her, either walking to the resident or helping 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 label five action classes (there are forty 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. Third, the ε-neighborhood graph is not densely connected in the nursing home dataset, as there are usually a few moving people in an image.

We can see that if we consider the graph structure as part of our model and 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 approaches with adaptive structures.

Comparison of the three proposed models: The structure-level approach and feature-level approach encode context in two different ways: from high-level inter-label dependencies and from low level feature descriptors. For the structure-level approach, our proposed learning algorithm is capable of selecting the useful context (person-person interaction) and ignoring the redundant. Experimental results demonstrate that the context selection strategy is very useful. The feature-level approach provides a flexible way to include context both spatially and temporally. It tends to include all the context in the neighborhood. Since the model does not have structures in the intermediate layer, this will not complicate inference. For some activities that do not have discriminative pairwise interactions (e.g. walking), a person's action usually benefits from knowing the dominant action of people nearby rather than a single person's interaction. In this case, the feature-level approach shows promising performance. One the other hand, for some activities such as talking and queuing, a pair of person's interaction (e.g. facing the same direction) is discriminative for the high level group activity. Thus selecting the context makes more sense than including everything. The combined approach makes a balance between the previous two approaches, and is thus more general for different group activities. Examples are in Fig. 7, where the structure-level approach shows the best performance for "queue", but the worst performance for "walk" compared to the other two approaches. The combined approach gives the best performance in terms of average accuracy.

6 Conclusion

In this paper, we have presented a novel framework for group activity recognition which jointly captures the group activity, the individual person actions, and the interactions among them. The goal of this paper is to demonstrate the effectiveness of contextual information in recognizing group activities. We have exploited two types of contextual information: *group-person interaction* and *person-person interaction*. In particular, we have proposed three different ways to model *person-person interaction*, one way is in the structure level, we have introduced an adaptive structures algorithm that automatically infers the optimal structure of person-person interaction in a latent variable framework. The second way is in the feature level, we have introduced an action context descriptor that encodes information about the action of an individual person in a video, as well as the behaviour of other people nearby. The third way combines the adaptive structure and the action context descriptor.

As future work, we would like to extend our model to consider multiple group activities in a scene at once. We also plan to investigate more complex structures, such as temporal dependencies among actions.

Acknowledgments

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR; grant numbers AMG-100487 and TIR-103945).

References

[1] I. Biederman, R. Mezzanotte, and J. Rabinowitz, "Scene perception: detecting and judging objects undergoing relational violations," *Cognitive Psychol.*, vol. 14, no. 2, pp. 143–177, 1982.

[2] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," in *NIPS*, 2010.

[3] T. Lan, Y. Wang, G. Mori, and S. Robinovitch, "Retrieving actions in group contexts," in *International Workshop on Sign Gesture Activity*, 2010.

[4] K. P. Murphy, A. Torralba, and W. T. Freeman, "Using the forest to see the trees: A graphicsl model relating features, objects, and scenes," in *NIPS*, 2004.

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

[6] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *ICCV*, 2007.

[7] A. Jain, A. Gupta, and L. S. Davis, "Learning what and how of contextual models for scene labeling," in *ECCV*, 2010.

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

[9] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *ICCV*, 2005.

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

[11] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*, 2009.

[12] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *ICCV*, 2009.

[13] H. Kjellstrom, J. Romero, D. M. Mercado, and D. Kragic, "Simultaneous visual recognition of manipulation actions and manipulated objects," in *ECCV*, 2008.

[14] R. Filipovych and E. Ribeiro, "Recognizing primitive interactions by exploring actor-object states," in *CVPR*, 2008.

[15] B. Yao and L. Fei-Fei, "Grouplet: a structured image representation for recognizing human and object interactions," in *CVPR*, 2010.

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

[17] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *PAMI*, 2009.

[18] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *CVPR*, 2010.

[19] T. Xiang and S. Gong, "Beyond tracking: Modelling activity and understanding behaviour," *IJCV*, 2006.

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

[21] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *CVPR*, 2004.

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

[23] W. Choi, K. Shahid, and S. Savarese, "What are they doing? : Collective activity classification using spatio-temporal relationship among people," in *Visual Surveillance*, 2009.

[24] N. Vaswani, A. Chowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," in *CVPR*, 2003.

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

[26] 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*, vol. 3, no. 8, pp. 509–520, 2006.

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

[28] S. S. Intille and A. Bobick, "Recognizing planned, multiperson action," *CVIU*, 2001.

[29] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *PAMI*, 2001.

[30] M. Ryoo and J. Aggarwal, "Stochastic representation and recognition of high-level group activities," *IJCV*, 2010.

[31] F. Cupillard, F. Bremond, and M. Thonnat, "Group behavior recognition with multiple cameras," in *CVPR*, 2002.

[32] M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu, "Group level activity recognition in crowded environments across multiple cameras," in *Workshop on Activity monitoring by multi-camera surveillance systems*, 2010.

[33] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2003.

[34] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *PAMI*, 2007.

[35] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition," in *CVPR*, 2009.

[36] C.-N. Yu and T. Joachims, "Learning structural SVMs with latent variables," in *ICML*, 2009.

[37] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[38] A. Vedaldi and A. Zisserman, "Structured output regression for detection with partial truncation," in *NIPS*, 2009.

[39] J. C. Niebles, C.-W. Chen, , and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV*, 2010.

[40] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *ECCV*, 2010.

[41] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *CVPR*, 2010.

[42] Y. Wang and G. Mori, "A discriminative latent model of image region and object tag correspondence," in *NIPS*, 2010.

[43] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection," in *CVPR*, 2005.

[44] T.-M.-T. Do and T. Artieres, "Large margin training for hidden markov models with partially observed states," in *ICML*, 2009.

[45] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

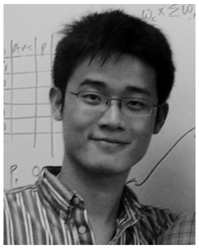
[46] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *PAMI*, 2000.

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

[48] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

Tian Lan is currently a Ph.D. candidate in the School of Computing Science at Simon Fraser University, Canada. He received his M.Sc. from the same university in 2010, and his B.Eng from Huazhong University of Science and Technology, China in 2008. He has worked as a research intern at Disney Research Pittsburgh in summer 2011. His research interests are in the area of computer vision, with a focus on semantic understanding of human actions and group activities within a scene.

Yang Wang is currently an NSERC postdoctoral fellow at the Department of Computer Science, University of Illinois at Urbana-Champaign. He received his Ph.D. from Simon Fraser University (Canada), his M.Sc. from University of Alberta (Canada), and his B.Sc. from Harbin Institute of Technology (China), all in computer science. He was a research intern at Microsoft Research Cambridge in summer 2006. His research interests lie in high-level recognition problems in computer vision, in particular, human activity recognition, human pose estimation, object/scene recognition, etc. He also works on various topics in statistical machine learning, including structured prediction, probabilistic graphical models, semi-supervised learning, etc.



Weilong Yang received the BSc degree in engineering from Southeast University, China, and the MSc degree in computer science from Simon Fraser University, Canada. He is currently a PhD candidate in school of computing science, Simon Fraser University. He interned in Google research in both summer and fall of 2010, and summer of 2011. His research interests are in human action recognition, even detection, and large-scale video tagging.

Stephen N. Robinovitch PhD (BAppSc-88, M.S.-90, Ph.D.-95), is a Professor and Canada Research Chair in Injury Prevention and Mobility Biomechanics at Simon Fraser University. His research focuses on improving our understanding of the cause and prevention of fall-related injuries (especially hip fracture) in older adults, through laboratory experiments, mathematical modeling, field studies in residential care facilities, and product design.

Greg Mori received the Ph.D. degree in Computer Science from the University of California, Berkeley in 2004. He received an Hon. B.Sc. in Computer Science and Mathematics with High Distinction from the University of Toronto in 1999. He is currently an associate professor in the School of Computing Science at Simon Fraser University. Dr. Mori's research interests are in computer vision, and include object recognition, human activity recognition, human body pose estimation.