Social Roles in Hierarchical Models for Human Activity Recognition

Tian Lan
Simon Fraser University
Burnaby, Canada
tla58@sfu.ca

Leonid Sigal
Disney Research
Pittsburgh, USA
lsigal@disneyresearch.com

Greg Mori
Simon Fraser University
Burnaby, Canada
mori@cs.sfu.ca

Abstract

We present a hierarchical model for human activity recognition in entire multi-person scenes. Our model describes human behaviour at multiple levels of detail, ranging from low-level actions through to high-level events. We also include a model of social roles, the expected behaviours of certain people, or groups of people, in a scene. The hierarchical model includes these varied representations, and various forms of interactions between people present in a scene. The model is trained in a discriminative max-margin framework. Experimental results demonstrate that this model can improve performance at all considered levels of detail, on two challenging datasets.

1. Introduction

In this paper we present a method for human activity recognition in entire scenes. Realistic scenes of human activity often involve multiple, inter-related actions at the same time. Levels of detail include low-level actions performed by individuals, groups of people involved in interactions, and an overarching scene level event. This work develops a model for integrating these varied levels of detail and sources of information that provide cues to infer the unknown labels at each level. We focus on the notion of *social roles*, modeling the expected behaviours of certain people, or groups of people, in a scene.

Consider the scene shown in Fig. 1. In terms of human activity recognition, there is a variety of questions one can ask. One can label this scene at multiple levels of detail. Who is the attacker? What are the players in the bottom right corner doing? How many people are running? Which players are defending (marking) members of the opposing team? What is the overall game situation? Note that potential queries often involve social roles such as "defender", "attacker" or "man-marking." In this paper we present a model towards answering queries such as these.

The representation of human activity is a challenging, open problem. It is arguable that multiple levels of detail and types of labels are required depending on the problem focus; we present a model that can be used to capture a variety of levels of detail in a unified framework. In addition to modeling of low-level actions (e.g. running or standing) and high-level events ("attack play", "penalty corner"), we model social roles. Social roles take into account inter-related people and are a complementary representation to the low-level actions typically used in the activity recognition literature. For example, a player engaged in the social role of "man-marking" is likely to have an opponent nearby.

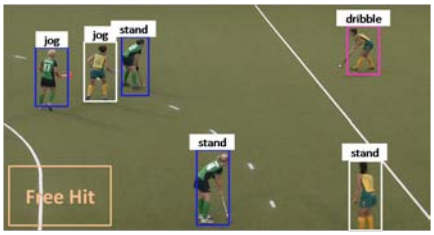


Figure 1. **An example of video event.** Beyond the general event description (e.g. Free Hit), we can explain the scene at multiple levels of detail such as low-level actions (e.g. standing and jogging) and mid-level social roles (e.g. attacker and first defenders). The social roles are denoted by different colors. In this example, we use magenta, blue and white to represent attacker, man-marking and players in the same team as the attacker respectively.

Further, the notions of low-level actions, social roles, and high-level events naturally require a contextual representation – the actions and social roles of all the people in a scene are interdependent, and related to the high-level event taking place. Our model captures these relationships, and allows flexible inference of the social roles and their dependencies in a given scene.

The main contribution of this paper is the development of this structured model of scenes of human activity. This model is hierarchical and contains representations ranging from low-level actions through social roles through to scene-level event class. We formalize parameter learning for this model in a max-margin framework. Moreover, this model can be used for a variety of human activity scene labeling tasks and queries. We demonstrate experimentally that it is effective, and that including social roles and rela-

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

tionships among them can result in improved performance on a variety of tasks.

2. Previous Work

Human activity recognition is an active area of research, see Poppe [20] for a survey. Much of the work focuses on recognition of low-level single-person actions (e.g. [22]). In this paper we rely on the low-level features and representations used by these methods to predict the actions of individuals, but build higher-level models upon them.

Previous work has also focused on human activity recognition in scenes, modeling interactions between individuals and higher-level group activities. Intille and Bobick [10] use probabilistic graphical models for recognizing hand-specified structured activities such as American football plays. Medioni et al. [15] reason about interactions between objects such as vehicles and road checkpoints. Moore and Essa [17] recognize multitasked activities. Cupillard et al. [5] presents an approach for recognizing specific activities such as violence or pickpocketing viewed by several cameras. Chang et al. [2] presents a real-time system to detect aggressive events in prison. Two hierarchical clustering approaches are proposed to group individuals, and events modeled 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. Ryoo and Aggarwal [21] propose a stochastic representation for more general group activities based on context-free grammars, which characterizes both spatial and temporal arrangements of group members. However, the representation of activities are encoded manually by human experts. Different from the aforementioned approaches, our work employs a structured SVM framework that is able to capture some structure of group activities, and the structures of group activities are learnt automatically. Patron-Perez et al. [19] also use a structured SVM framework, with a focus on activities that are defined on a pair of interacting humans. Gupta et al. [9] use AND-OR graphs to represent complex events.

Another important cue for disambiguating actions is the context provided by the actions of nearby humans in the same scene. A few recent approaches include a model of group activity context, e.g., Choi et al. [3, 4], Lan et al. [13], and Amer and Todorovic [1]. We build on this line of work, but include person-person interactions as context at varying levels of detail and representation.

Our model also incorporates high-level information about the overall event present in a scene. There has also been much effort on scene-level representations of activity, much of it using unsupervised learning. Loy et al. [14] model regions of activity and their relationships over surveillance videos. Wang et al. [24] examine hierarchical representations, and apply them to traffic scenes. Mehran et al. [16] aim to discover anomalous events in surveillance video by analyzing low-level motion cues. Kuettel et al. [12] build temporal latent topic models that can be used to discover patterns of activity in traffic surveillance video. These models have a similar aim to ours in terms of modeling high-level activities and relationships between entities, but typically work at a lower feature level of detail and don't explicitly model interactions.

Most prior methods to activity recognition focus on one of these levels of abstraction (or detail), using others, typically lower levels, as latent intermediate representations of little inherent interest. In contrast, our model is explicitly designed to support semantic level-of-detail inferences in the context of rich multi-person event scenes.

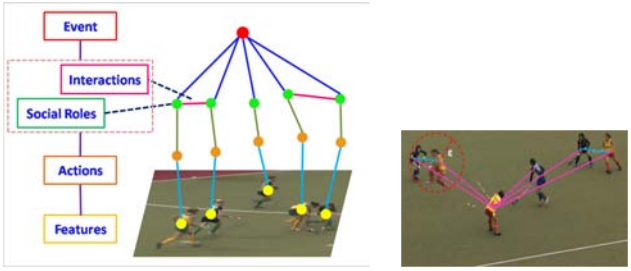
3. Modeling Video Event Structures

In this paper we develop a model for human activity in an entire scene. The model is hierarchical, and includes various levels of detail: low-level actions, mid-level social roles, and high-level events. The relationships and interactions between these are included in the model.

This model is general, and we show two different application domains in our experiments. However, to ground the model description, we describe an instantiation applicable to modeling field hockey videos. We define 11 low-level action classes: pass, dribble, shot, receive, tackle, prepare, stand, jog, run, walk, save; 5 social roles: attacker, first defenders, defenders defend against person (man-marking), defenders defend against space, other; and 3 scene-level events: attack play, free hit and penalty corner. Each person is labeled with both action and social role. Attacker is the person who controls the ball, defenders are classified into three categories: first defenders directly defend against the attacker, other defenders either defend against person or space. Players from the same team of the attacker, the referee and the goalie are assigned the label of "other".

We propose a discriminative model for learning the hierarchical structures of video events in this domain. The model is a general framework that can carry out different inferences based on a user's preference (e.g., finding the attacker, action recognition and social role recognition of each player). This is done by modifying the learning criterion (i.e., the loss function) while keeping the model structure unchanged.

A graphical illustration of the model is shown in Fig. 2. At the lowest level, the compatibility between a person's feature vector (e.g., HOG [6]) and action is modeled. At the intermediate level, the model explores the contextual information by modeling interactions between people in terms of their social roles. Social roles naturally capture important interactions, e.g. *first defenders* tend to appear in the neighborhood of an *attacker*, *man-marking* happens when there is a player from the opposing team. On the top level



(a) (b)

Figure 2. **Graphical illustration of the model.** Different types of potentials are denoted by lines with different colors in (a). An example of the graph structure $\mathcal{G}$ of the intermediate layer is in (b): an attacker is connected to every other player, non-attackers are connected to the closest player within a distance of ε.

of the model are the scene-level events. Events are inter-dependent with social roles, e.g., during a penalty corner, *attacker* appears in the corner of the field.

3.1. Model Formulation

We first describe the labeling. We assume an image has been pre-processed, so the location of the goal and persons in the image have been found. We separate the players into two teams according to their color histograms. Each person is associated with two labels: action and social role. Let $h_i \in \mathcal{H}$ and $r_i \in \mathcal{R}$ be the action and social roles of the person i respectively, where $\mathcal{H}$ and $\mathcal{R}$ are the sets of all possible action and social role labels respectively. Each video sequence is associated with an event label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all possible event labels.

We define the score of interpreting a video sequence I with the hierarchical event representation as:

$$F_w(\mathbf{x}, y, \mathbf{r}, \mathbf{h}, I) = w^\top \Phi(\mathbf{x}, y, \mathbf{r}, \mathbf{h}, I) = \sum_j w_1^\top \phi_1(x_j, h_j) + \sum_j w_2^\top \phi_2(h_j, r_j) + \sum_{j,k} w_3^\top \phi_3(y, r_j, r_k) \quad (1)$$

Action model $w_1^\top \phi_1(x_j, h_j)$: This potential function is a standard linear model trained to predict the action label of the j-th person. 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 $\mathbb{1}(\cdot)$ is the indicator function. x_j is the feature vector extracted from the j-th person. In order to encode temporal information, we extract 3-frame tracklets for each person in a video based on data association, and x_j is computed by concatenating the HOG [1] extracted from the person bounding box in three consecutive frames.

[1]We use the code available at http://www.cs.brown.edu/ pff/latent/ for computing the HOG descriptor.

Unary role model $w_2^\top \phi_2(h_j, r_j)$: This potential function represents the properties of social roles, including dependencies between the action label h_j and social role r_j, and the role-specific locations of the j-th person. It is parameterized as:

$$w_2^\top \phi_2(h_j, r_j, I) = \sum_{c \in \mathcal{R}} \sum_{b \in \mathcal{H}} w_{2cb} \cdot \mathbb{1}(h_j = b) \cdot \mathbb{1}(r_j = c) + \sum_{c \in \mathcal{R}} \sum_{m \in \mathcal{M}} w_{2ck} \cdot \mathbb{1}(r_j = c) \cdot \text{bin}_m(j) \quad (3)$$

Here we divide an image into M cells, $\text{bin}_m(j) = 1$ if the j-th person falls into the m-th cell, otherwise 0. The spatial binning is determined with the goal's location as a reference frame. For example, if the goal is on the left of the image, then the first bin starts at the bottom left of the image.

Pairwise role model $w_3^\top \phi_3(y, r_j, r_k)$: This potential function represents the dependencies between of a pair of social roles r_j and r_k under an event y. It is parameterized as:

$$w_3^\top \phi_3(y, r_j, r_k) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{R}} \sum_{c \in \mathcal{R}} w_{3abc}^\top \mathbb{1}(y = a) \cdot \mathbb{1}(r_j = b) \cdot \mathbb{1}(r_k = c) \cdot d_{jk} + \sum_{b \in \mathcal{R}} \sum_{c \in \mathcal{R}} w_{3bc}^\top \mathbb{1}(r_j = b) \cdot \mathbb{1}(r_k = c) \cdot g_{jk} \quad (4)$$

We use a similar spatial context feature as in [7]: d_{jk} is the feature that bins the relative location of the j-th and k-th person into one of D spatial relations: overlap, next-to, near, above and below. It is a sparse vector of all zeros with a single one for the bin occupied by the satisfied spatial relation. Similar to the *unary role model*, we use the goal's location in an image as a reference. For example, if the goal appears at the right of an image, then d_{jk} is "above" when j-th person is to the right of k-th person. g_{jk} is a two-dimensional vector: $g_{jk} = [1, 0]$ if j and k are in the same team, and $g_{jk} = [0, 1]$ otherwise.

We use an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the mid-level social roles and interactions, where a vertex $v_i \in \mathcal{V}$ corresponds to the social role r_i, and an edge $(v_j, v_k) \in \mathcal{E}$ corresponds to the interactions between r_j and r_k. In our case, the attacker is connected to every other player, the non-attackers are connected to the closest player within a distance of ε. In the case that there is no attacker in a video frame, then every player is connected to the closest player within a distance of ε. An example graph structure is shown in Fig. 2(b).

4. Learning

We assume we are given a set of training examples with the actions, social roles, and event labels. Given a set of N such training examples $\langle \mathbf{x}^n, y^n, \mathbf{h}^n, \mathbf{r}^n, I^n \rangle$ $(n = 1, 2, \ldots, N)$, we would like to train the model parameter $\mathbf{w}$

that tends to produce the correct hierarchical event structures that include event y, social roles $\mathbf{r}$ and actions $\mathbf{h}$ given a new test video. A natural way of learning the model is to adopt the structured SVM formulation [11] as follows:

$$\min_{w,\xi\geq 0} \frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\xi_n$$
$$\text{s.t. } F_w(\mathbf{x}^n, y^n, \mathbf{r}^n, \mathbf{h}^n, I^n) - F_w(\mathbf{x}^n, y, \mathbf{r}, \mathbf{h}, I^n) \geq$$
$$\Delta(y, y^n, \mathbf{h}, \mathbf{h}^n, \mathbf{r}, \mathbf{r}^n) - \xi_n, \forall n, y, \mathbf{h}, \mathbf{r} \quad (5)$$

where $\Delta(y, y^n, \mathbf{r}, \mathbf{r}^n, \mathbf{h}, \mathbf{h}^n)$ measures the joint loss between the ground-truth event label, social role labels and action labels $(y^n, \mathbf{r}^n, \mathbf{h}^n)$ compared with the hypothesized ones $(y, \mathbf{r}, \mathbf{h})$. We define the joint loss as a weighted combination of the loss on different terms $\Delta(\mathbf{y}, \mathbf{y}^n, \mathbf{r}, \mathbf{r}^n, \mathbf{h}, \mathbf{h}^n) = \Delta_{0/1}(y, y^n) + \nu\sum_i \Delta_{0/1}(r_i, r_i^n) + (1-\mu-\nu)\sum_i \Delta_{0/1}(h_i, h_i^n))$, where $0 \leq \mu \leq 1, 0 \leq \nu \leq 1$ balance the contribution of terms.

This is a rather general learning framework that can carry out different inferences based on a user's preference (e.g., finding the attacker, action and social role recognition for each player). This is done by modifying the learning criterion (i.e., the loss function) while keeping the model structure unchanged. For example, if the user's preference is social role recognition, then we can simply set μ to zero to make the formulation directly optimize the social roles .

The main computational challenge in structured SVM learning is *loss augmented inference* or finding the *most violated constraint*. This is a special case of the general inference algorithm described next.

5. Inference

Given a test video there is a variety of queries one might wish to answer. Using our hierarchical model, one can formulate queries about any individual variable at any level of detail. For instance, one can query on the overall event label y for the scene, or the social role label r_j of a particular person.

We examine the margin, or difference in model scores between values, for an individual variable to give a score for its setting. Given a video and a query variable q, the inference problem is to find the best hierarchical event representation that maximizes the scoring function $F_w(\mathbf{x}, y, \mathbf{h}, \mathbf{r}, I)$ while fixing the value of q to its possible values. For example, if q is the action of one person (one of the h_i), we would compute the maximum value of the scoring function F_w when fixing q to each possible action. We then set the score for the person to be performing each action as the difference between this score and that of the next best action.

For a given video and query variable q, this inference requires solving the following optimization problem:

$$\max_{y,\mathbf{h},\mathbf{r}\setminus q} F_w(\mathbf{x}, y, \mathbf{h}, \mathbf{r}, I) = \max_{y,\mathbf{h},\mathbf{r}\setminus q} w^\top \Phi(\mathbf{x}, y, \mathbf{h}, \mathbf{r}, I) \quad (6)$$

The optimization problem in Eq. 6 is in general NP-hard since it involves a combinatorial search. We instead use a coordinate ascent style algorithm to approximately solve Eq. 6 by optimizing one variable at a time while fixing the other two variables, these steps iterate until convergence. Since the action labels do not have any structures, we can simply enumerate all the possible $h \in \mathcal{H}$ to predict the the best action label h^*. We use the same strategy for predicting the best event label y^*. Optimizing the social roles $\mathbf{r}$ is more challenging, since the graph structure $\mathcal{G}$ and $\mathbf{r}$ are correlated. In the following, we develop methods for optimizing the social roles $\mathbf{r}$.

During inference, the graph structure $\mathcal{G}$ depends on which person is the attacker. Once the attacker is fixed, then the graph structure is fixed. Suppose there are K people in a video clip, we enumerate all possible situations $(k = 0, 1, \ldots, K)$, each person k is regarded as an attacker one at a time if $k \geq 1$; we also consider the case that there is no attacker in the video clip ($k = 0$). This is equivalent to enumerating all possible graph structures.

We introduce variables $\mathbf{r}^k$ to denote the social roles $\mathbf{r}$ when person k is the attacker, i.e. $r_k = 1$. The inference of $\mathbf{r}$ requires solving the following optimization problem:

$$\max_{0\leq k\leq K} \max_{\mathbf{r}^k} w^\top \Phi(\mathbf{x}, y, \mathbf{h}, \mathbf{r}^k, I) \quad (7)$$

The inner maximization of $\mathbf{r}^k$ with a fixed k is a standard max-inference problem in an undirected graphical model. Here we use loopy BP to approximately solve it.

6. Experiments

In order to demonstrate our proposed method, we consider two different experimental scenarios. First, we test our approach on highly structured activities. We present a new challenging *Broadcast Field Hockey Dataset* that consists of sporting activities captured from broadcast cameras. Second, we consider more general activities in daily living. We test our model on a dataset of surveillance videos recorded in a nursing home [13]. In order to show that our model can perform various inferences based on a user's preference, we test the model with different tasks including action and social role recognition, searching for specific social roles and scene-level events.

The focus of this paper is activity recognition. In order to test our model for activity, we perform experiments using ground-truth person locations, as well as those using a simple automated detector. Automatic detection and tracking, particularly in sports and surveillance videos, is an established area of research. Many excellent methods (e.g. [18, 23]) automatically generate high-quality person locations similar to our ground-truth results.

We compare our model with several baseline methods. The first baseline is the action model in Eq. 2. It is equiv-

alent to an SVM with linear kernel based on the feature vector of each person. This baseline can only perform action recognition. The second baseline (which we call *unary*) consists of both the action model (Eq. 2) and the unary role model (Eq. 3). In order to compare with our model for event recognition, we further add links between the event $\mathbf{y}$ and social roles $\mathbf{r}$ with the potential function: $\sum_j \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{R}} w_{ab}^\top \mathbb{1}(y = a) \cdot \mathbb{1}(r_j = b)$. We also re-implement the adaptive structured latent SVM method in [13] and compare to it on the *Nursing Home Dataset* [13].

In our datasets, the action/social role classes are extremely imbalanced (e.g. the number of attackers is less than 1/10 of all the examples). In this case, the traditional 0-1 loss $\Delta_{0/1}$ defined in Eq. 5 is not appropriate. We adopt a margin rescaling approach by using the loss function introduced in [25] to handle this problem:

$$\Delta_{bal}(u, u^n) = \begin{cases} \frac{1}{m_p} & \text{if } u \neq u^n \text{ and } u^n = p \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where m_p is the number of examples with class label p.

Rather than directly using certain raw features (e.g. the HOG descriptor [6]) as the feature vector x_i in our framework, we train a multi-class SVM classifier based on the raw feature of each individual and their associated action labels. In the end, each feature vector x_i is represented as an N-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.

6.1. Broadcast Field Hockey Dataset

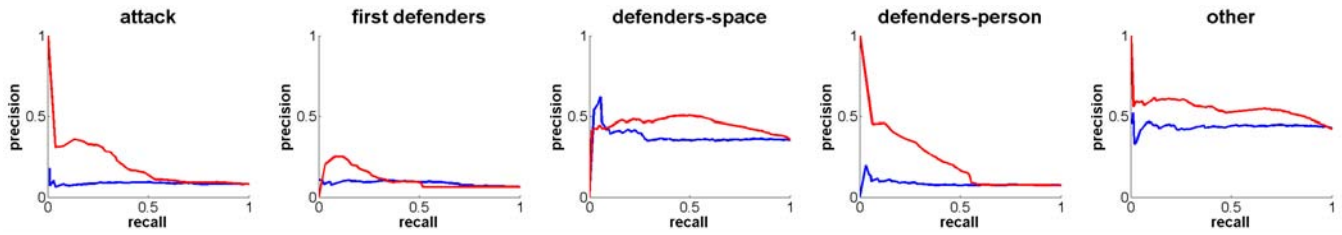
We have collected a new challenging dataset of broadcast field hockey games. Our dataset contains 58 video sequences extracted from five matches. These video sequences are highlights of field hockey games with 11 action classes, 5 social role classes and 3 scene-level event classes. See Fig. 3 for example frames from the dataset. The human activities depicted in the dataset contain complex person-person interactions that go beyond simple actions and group activities. For instance, sequences from a typical attack play usually include interactions such as passing the ball, dribbling and tackling, receiving the ball and defending, etc. We use Leave-One-Out (LOO) cross validation in our experiments, cycling each video sequence as a test video one at a time.

Results: We summarize the mean per-class accuracies for action, social role and scene-level event recognition in Table 1. We can see that our model significantly outperforms the baselines in terms of all of the three tasks. In terms of social role recognition, our model provides a strong improvement of more than 20% over the baseline. The comparison of confusion matrices, for social role recognition, between our method and the baseline *unary* are illustrated in Fig. 4. As one can see, the task is almost unachievable by only using the unary term of social roles (the accuracy is close to chance). However, the confusions between different social roles are significantly reduced by including the pairwise relations between players. We also report results in Table 2 using the LSVM person detector [8] trained on our dataset. We could draw similar conclusions as using the ground truth person locations. Note that we treat the false positives of the detector as incorrect predictions in reporting the classification accuracies. The average precision of person detection on this dataset using the LSVM detector is 33.67.

Figure 3. **Broadcast Field Hockey Dataset.** Our dataset contains 58 video sequences with 11 actions: pass, dribble, shot, receive, tackle, prepare, stand, jog, run, walk, save; 5 social roles: attacker, first defenders, defenders defend against person, defenders defend against space, other; 3 scene-level events: attack play, free hit, and penalty corner, images in each row denote an event class. These video sequences are highlights extracted from five field hockey matches recorded in the Dublin Stadium by broadcast cameras.

The benefits of including the pairwise social roles can be further demonstrated by the learned pairwise weights, as visualized in Fig. 6. Our model learns meaningful multi-class spatial layouts and team memberships – e.g., attacker and defenders tend to be from opposing teams, defenders defend against space and person tend to appear above the attacker. For the low level action recognition and high-level event recognition, our model again provides a noticeable improvement over the baselines. The mid-level social roles help by modeling which actions tend to be performed by people in which social roles. Further, the structured interactions between the people in these social roles can be indicative of certain high-level events. We believe this is the reason for the improved performance in low-level action and high-level event recognition when using social roles.

Fig. 5 shows the comparison of precision-recall curves for inference of different social roles using our method and the baseline. Our model outperforms the baseline by significant margins in all of the five social roles. Fig. 7 shows the visualizations of our predicted events and social roles.



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

Figure 5. **Precision recall curves.** Our method in (red) and the baseline in (blue) applied in Broadcast Field Hockey Dataset to the task of detecting: (a) attacker (b) first defenders (c) defenders defend against space (d) defenders defend against person (e) other.

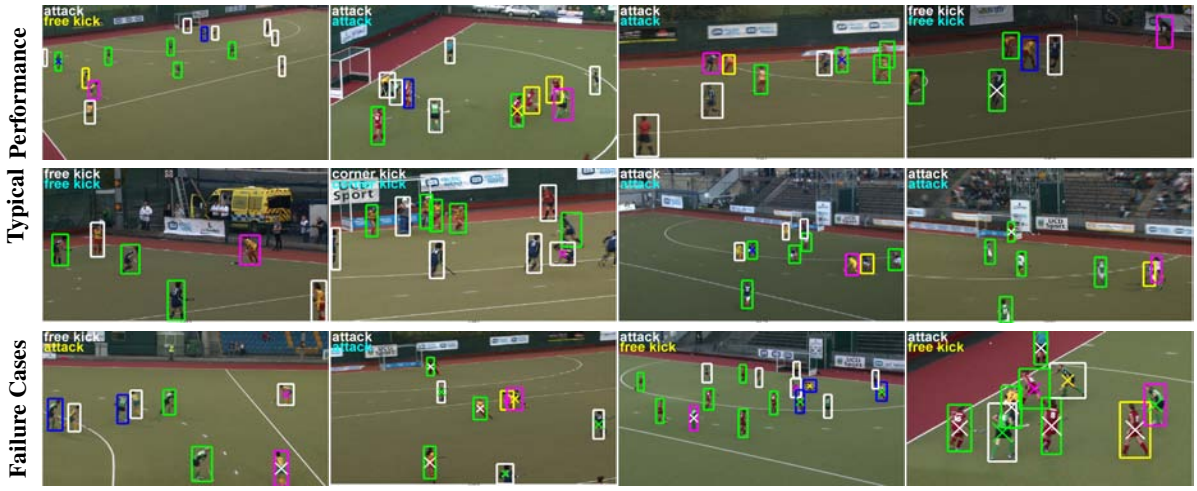
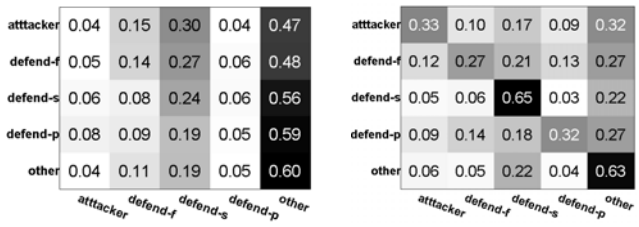


Figure 7. **Visualization of our results on Broadcast Field Hockey Dataset.** The ground truth event (white) and the predicted event are shown in the left corner of each image. Correct predictions are visualized in blue, otherwise yellow. Each bounding box is represented by a color, which denotes the predicted social roles. We use magenta, yellow, green, blue and white to represent the social roles attacker, first defenders, defenders defend against space, defenders defend against person and other respectively. The cross sign in the middle of a bounding box indicates incorrect predictions, and the ground truth social roles are indicated by the color of the cross sign. The last row shows bad predictions, all of them result from incorrect prediction of the attacker, since the social roles of other players are highly dependent on the attacker.



(a) (b)

Figure 4. **Confusion matrices for social role recognition.** Illustrated are accuracies on the Broadcast Field Hockey Dataset: (a) unary (stands for the unary term of social roles) (b) our approach. Rows are ground-truths, and columns are predictions. Each row is normalized to sum to 1. Here we use defend-f, defend-s, defend-p to represent first defenders, defenders defend against space and defenders defend against person respectively.

Method	Role	Event	Action
unary	21.7	56.9	21.5
full model	**44.0**	**62.9**	**28.8**
action model	N/A	N/A	26.1

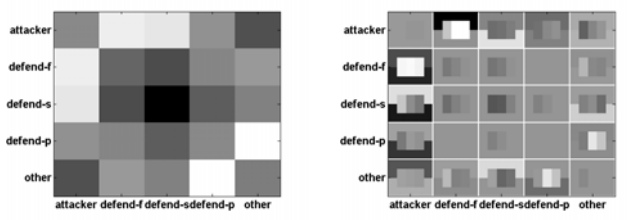
Table 1. Comparison of social role, event and action classification accuracies (Mean per-class) of different methods on the Broadcast Field Hockey Dataset.

Method	Role	Event	Action
unary	18.9	48.9	15.3
full model	**27.0**	**50.6**	**17.7**
action model	N/A	N/A	17.2

Table 2. Comparison of social role, event and action classification accuracies (Mean per-class) of different methods on the Broadcast Field Hockey Dataset with automated person detector.

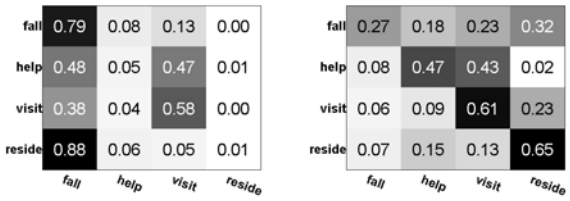
6.2. Nursing Home Dataset

The second dataset [13] 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, and the scene-level events include fall and non-fall. We further define four social roles: *fall,*



(a) (b)

Figure 6. **Visualization of the learnt pairwise weight.** In (a) are opposing teams and in (b) spatial relations, under the event *attack play*. Light cells indicate large values of weights. Consider the example (a), the model favors seeing first defenders and the attacker from opposing teams, defenders defend against person and "other" (usually refer to the players being defended) from opposing teams are also favored. (b) the pairwise spatial weights for each pair of classes are represented as an image patch: the three bins in the middle of each patch denote the spatial relations: near, next-to, overlap; the other two bins in the surroundings of each patch represent the spatial relations: above and below. We can see the attacker and first defenders tend to be close to each other, and the attacker tend to appear below the defenders defend against space.



(a) (b)

Figure 8. **Confusion matrices for social role recognition.** Illustrated are accuracies on the Nursing Home Dataset using: (a) unary and (b) our approach. Rows are ground-truths, and columns are predictions. Each row is normalized to sum to 1.

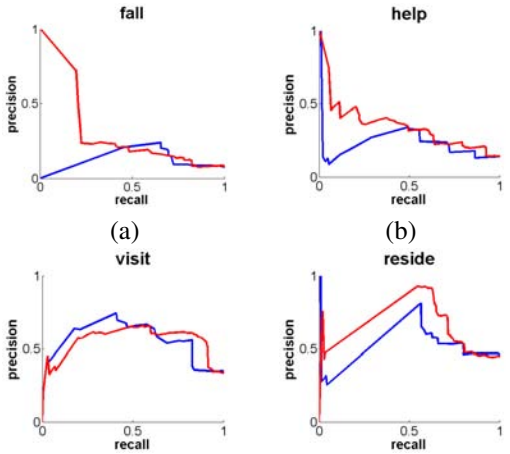
help, visit and reside. The fallen person is labeled as "fall", the people helping the fallen person are labeled as "help", "visit" happens when the nurses come to the nursing home to talk to the residents or clean the room, the residents sitting or standing in the nursing home are labeled as "reside". We use the same training/testing splits as [13].

The dataset is extremely challenging because of low framerate and spatial resolution, which makes it difficult to detect falls based on the features of a single person. Social roles can be very helpful in this application. For example, when a person falls, typically people come to help and thus interact with the person lying on the ground.

Results: We summarize the mean per-class accuracies for action, social role and scene-level event recognition in Table 3. We can see that our model significantly outperforms the other baselines in terms of all of the three tasks. In order to show the contribution of social roles in event recognition, we re-implement the adaptive structured latent SVM

Method	Role	Event	Action
unary	35.0	73.2	40.9
full model	**50.1**	**80.5**	**42.0**
action model	N/A	N/A	38.7
Lan et al. [13]	N/A	78.5	N/A

Table 3. Comparison of social role, event and action recognition accuracies (Mean per-class) of different methods on the Nursing Home Dataset.



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

Figure 9. **Precision recall curves.** Our method in (red) and the baseline in (blue) applied in the Nursing Home Dataset to the task of detecting: (a) fall, (b) help, (c) visit and (d) reside.

method proposed in [13], the comparison shows the improvement of our method. We further ran a two-tailed T-test and verified that the improved performance with respect to [13] is statistically significant at alpha=0.05 (p-value = 0.75).

Fig. 8 shows the confusion matrices of our method and the baseline *unary* in terms of social role recognition. Similar to the first dataset, the baseline almost fails at separating the social roles. The pairwise model provides a strong improvement in reducing the confusions between different social roles.

Fig. 9 shows the comparison of precision recall curves for searching different social roles of our method and the baseline. We can see that our method significantly outperforms the baseline in three social roles: *fall, help* and *reside*. For *visit*, our method performs similarly to the baseline, this is because searching for *visit* rarely requires contextual information, since the nurse usually visits the nursing home alone. In this case, pairwise relations would not help. Fig. 10 shows the visualizations of our predicted events and social roles.

7. Conclusion

We have developed a structured model for human activity recognition in complex scenes. The model integrates a

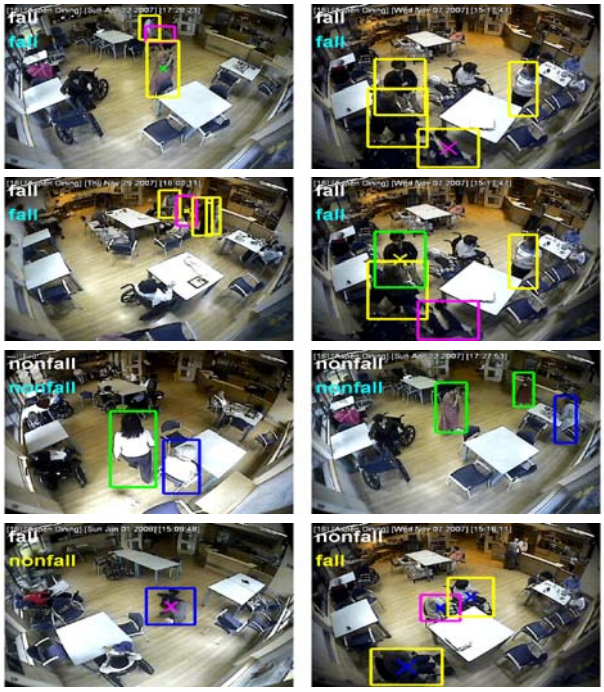


Figure 10. **Visualization of our results on the Nursing Home Dataset.** The ground truth event (white) and the predicted event are shown in the left corner of each image. Please refer to Fig. 7 for the visualization rules. We use magenta, yellow, green and blue to represent the social roles fall, help, visit and reside respectively. The first three rows show the examples where the pairwise relations help predict falls and non-falls. The last row shows incorrect predictions.

variety of levels of detail including the low-level actions, mid-level social roles, and high-level events. In particular, we have presented a new representation – social roles for human activity recognition as a complementary representation to the typical low-level actions. The advantage of this model is that it naturally captures the interdependencies between actions, social roles and high-level events, and allows flexible inference of the social roles and their dependencies in a given scene. The model parameters are learned in a max-margin framework. Our experimental results demonstrate that our model is effective in performing a variety of inference tasks including action, social role and event recognition. We illustrate that including social roles in the model results in improved performance on all of these tasks.

References

[1] M. R. Amer and S. Todorovic. A chains model for localizing participants of group activities in videos. In *ICCV*, 2011.

[2] M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu. Group level activity recognition in crowded environments across multiple cameras. In *AMMCSS*, 2010.

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

[4] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011.

[5] F. Cupillard, F. Bremond, and M. Thonnat. Group behavior recognition with multiple cameras. In *WACV*, 2002.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

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

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

[9] 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.

[10] S. S. Intille and A. Bobick. Recognizing planned, multiperson action. *CVIU*, 81:414–445, 2001.

[11] T. Joachims. Training linear SVMs in linear time. In *SIGKDD*, 2006.

[12] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010.

[13] T. Lan, Y. Wang, W. Yang, S. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *PAMI*, 2011.

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

[15] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *PAMI*, 23(8):873–889, August 2001.

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

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

[18] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking – linking identities using bayesian network inference. In *Proc. CVPR*, 2006.

[19] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in tv shows. In *BMVC*, 2010.

[20] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28:976–990, 2010.

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

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

[23] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua. Tracking multiple people under global appearance constraints. In *ICCV*, 2011.

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

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