SFU

Social Roles in Hierarchical Models for Human Activity Recognition

Tian Lan[†] Leonid Sigal[‡] Greg Mori[†]

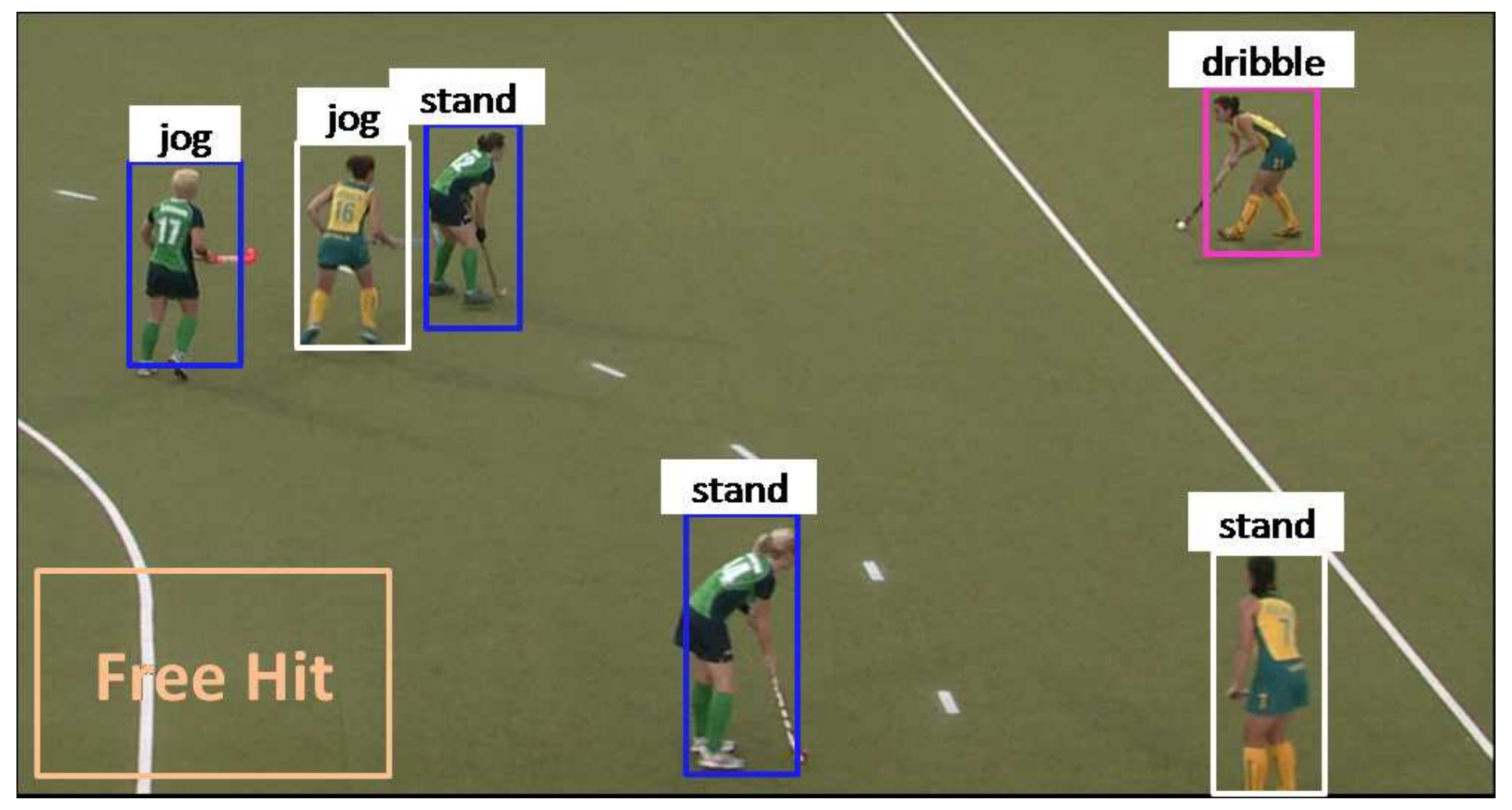
[†]Simon Fraser University [‡]Disney Research Pittsburgh

Disney Research

Overview

Problem:

- Realistic scenes of human activity often involve multiple, inter-related actions at the same time.
- A variety of questions one can ask for a scene (e.g. a hockey game): Who is the attacker? How many people are running? What is the overall game situation? We present a model towards answering queries such as these.

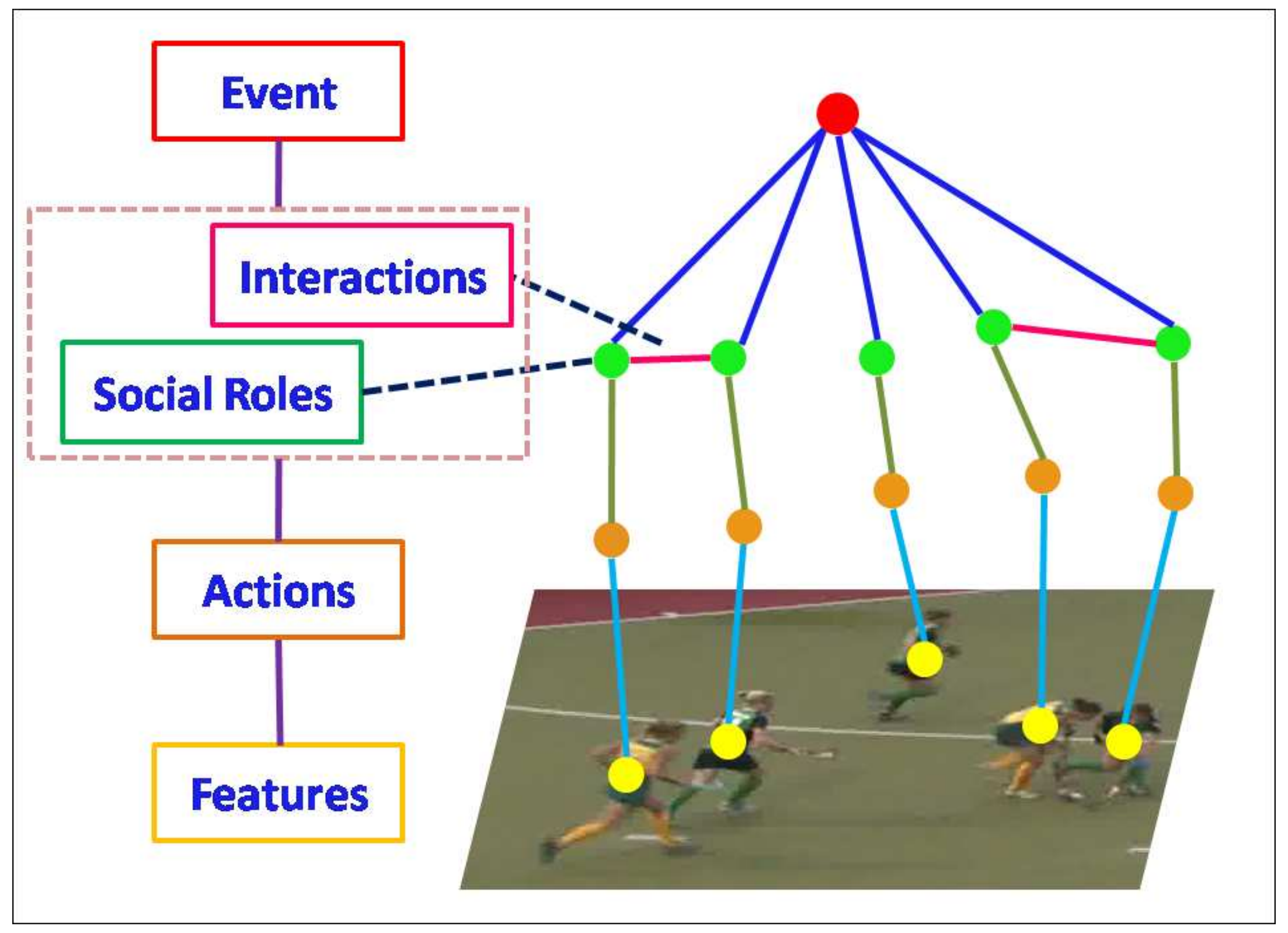


Our contributions:

- A new representation of human activity in multiple levels of detail, ranging from low-level actions through social roles through to scene-level event class.
- A hierarchical model could answer various queries and label the scene in a unified framework.
- Social roles, modeling the expected behaviours of certain people, or groups of people, in a scene.
- A new challenging Broadcast Field Hockey Dataset is collected.

Modeling Structures of Human Activities

Graphical Representation:



Event

Interactions

Social Roles

Actions

Features

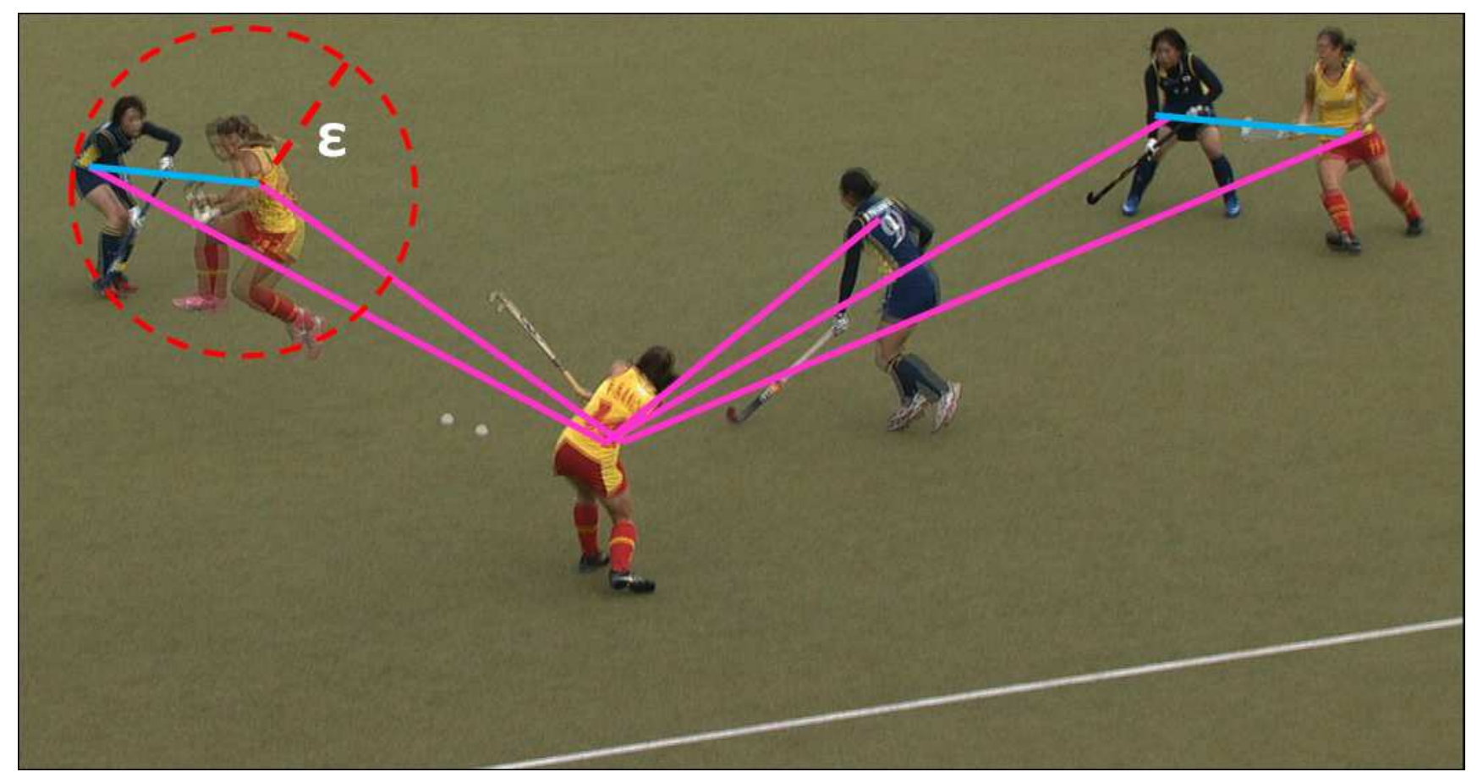
- The model includes various levels of detail: low-level actions, mid-level social roles, and high-level events.
- At the intermediate level, the model explores interactions between people in terms of their social roles.

Model Formulation

- Scoring function for a video with image features $\mathbf{x}$, action labels $\mathbf{h}$, social roles $\mathbf{r}$ and scene-level event label y:

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

- Action model: $w_1^\top \phi_1(x_j, h_j)$
  - a standard linear model trained to predict the action label of a person
- Unary role model: $w_2^\top \phi_2(h_j, r_j)$
  - action - social role dependencies
- Pairwise role model: $w_3^\top \phi_3(y, r_j, r_k)$
  - dependencies between a pair of social roles under an event
- Graph structure of social roles



Learning and Inference

Max-margin Learning

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

- A joint loss on event, social roles and actions.

$$\Delta\left(y, y^n, \mathbf{r}, \mathbf{r}^n, \mathbf{h}, \mathbf{h}^n\right) = \Delta_{0/1}\left(y, y^n\right) + \nu \Delta_{0/1}\left(\mathbf{r}, \mathbf{r}^n\right) + (1 - \mu - \nu)\Delta_{0/1}\left(\mathbf{h}, \mathbf{h}^n\right))$$

- A general learning framework that can carry out different inferences based on a user's preference

Inference

- One can formulate queries about any individual variable at any level of detail.
- For a given video and query variable q, the inference is to find the best hierarchical event representation while fixing the query q to its possible values.

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

- Approximate inference: optimize one variable at a time while fixing the other two variables, iterate until convergence.

Experiments

Datasets:

- Broadcast Field Hockey Dataset
  - We collected this new challenging dataset for human activity recognition.
  - The videos are highlights from five **real** field hockey matches.
  - Typical social roles: attacker, first defender, man-marking, etc.
- Nursing Home Dataset
  - Recorded in a dining room of a nursing home.
  - Typical social roles: fall, visit, reside, help.

Recognition:

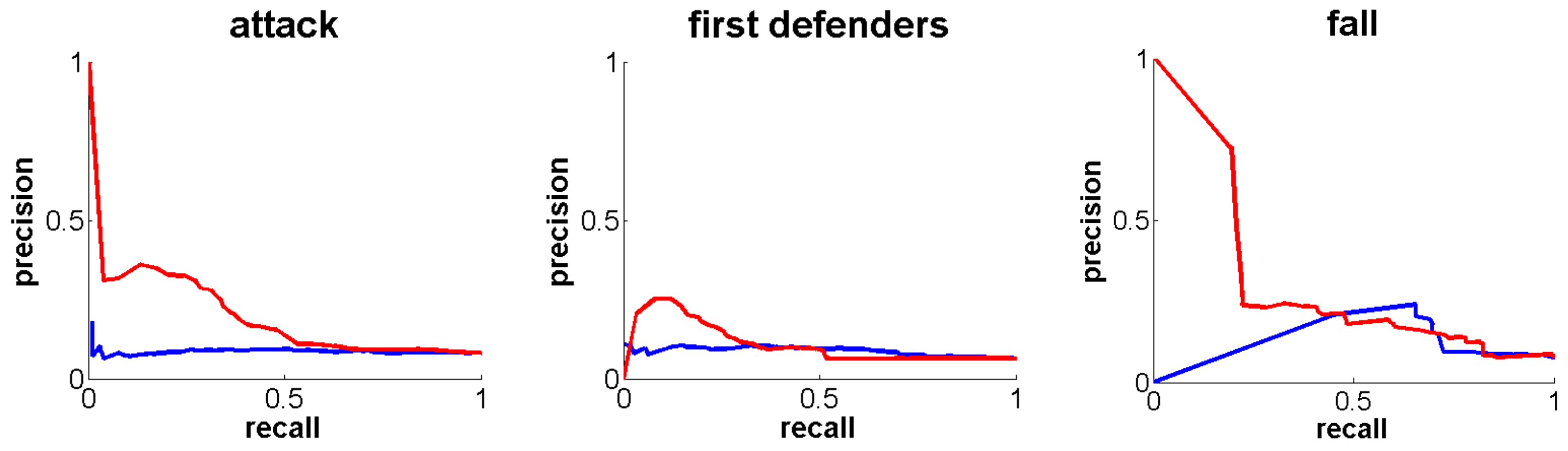
Method	Role	Event	Action
unary	21.7	56.9	21.5
full model	**44.0**	**62.9**	**28.8**
action model	N/A	N/A	26.1

broadcast field hockey dataset

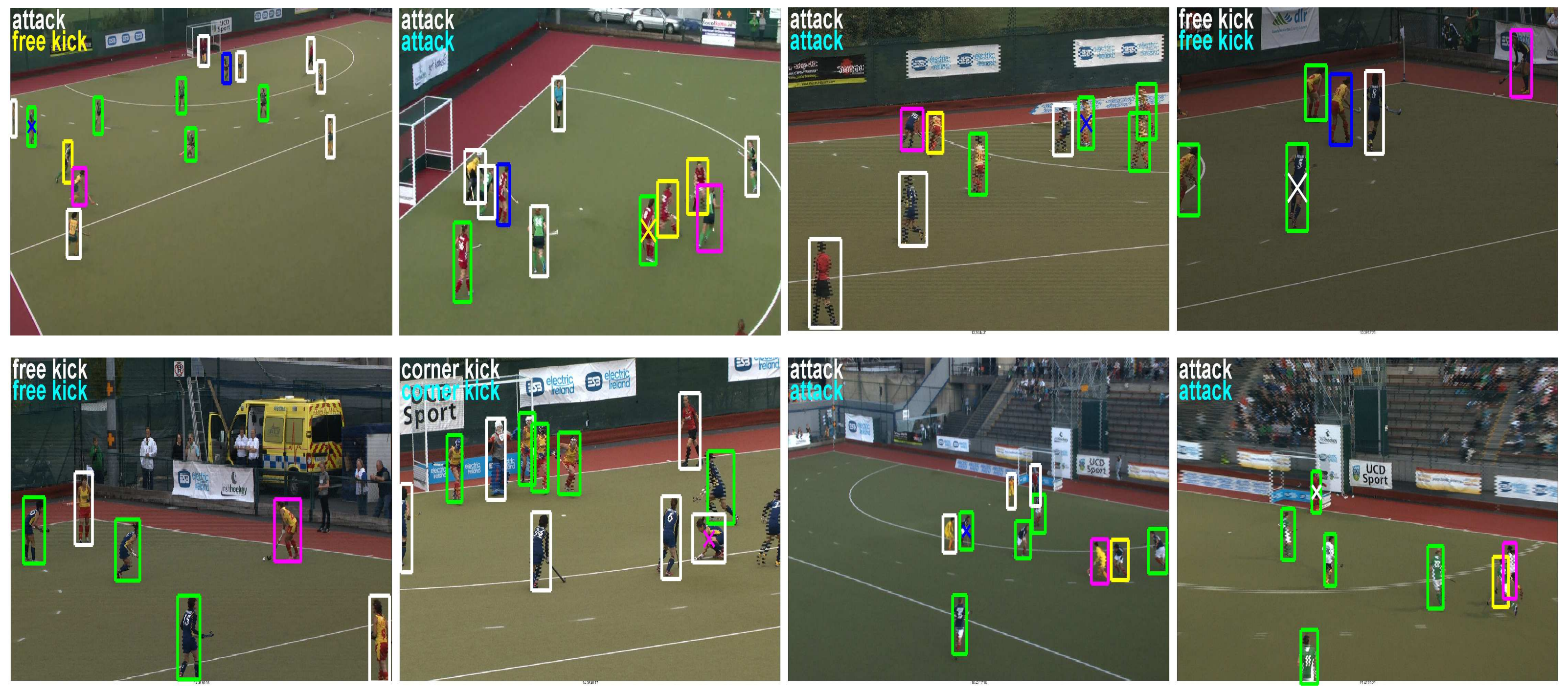
Method	Role	Event	Action
unary	35.0	73.2	40.9
full model	**50.1**	**80.5**	**42.0**
action model	N/A	N/A	38.7
Lan et al.	N/A	78.5	N/A

nursing home dataset

Searching for specific social roles:



full model (red) vs. unary role model (blue)



attacker (magenta), first defenders (yellow) defend against space (green), defend against person (blue), other (white).

fall (magenta), help (yellow), visit (green), reside (blue).

IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, June 2012