This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Classification of Puck Possession Events in Ice Hockey

Moumita Roy Tora Jianhui Chen James J. Little
Department of Computer Science
University of British Columbia
{moumita, jhchen14, little}@cs.ubc.ca

Abstract

Group activity recognition in sports is often challenging due to the complex dynamics and interaction among the players. In this paper, we propose a recurrent neural network to classify puck possession events in ice hockey. Our method extracts features from the whole frame and appearances of the players using a pre-trained convolutional neural network. In this way, our model captures the context information, individual attributes and interaction among the players. Our model requires only the player positions on the image and does not need any explicit annotations for the individual actions or player trajectories, greatly simplifying the input of the system. We evaluate our model on a new Ice Hockey Dataset. Experimental results show that our model produces competitive results on this challenging dataset with much simpler inputs compared with the previous work.

1. Introduction

Computer vision has been widely used in many sports applications [22]. The applications are expanded from information extraction such as player detection/tracking [20] to new visual information generation such as free-viewpoint video generation [11], and further to prediction of shot location [32] and broadcast camera angle planning [5].

Among various applications, group activity recognition is an active research area. Group activity recognition refers to determining what a group of people are doing, providing semantic and abstract descriptions for a sequence of images. Sports activity recognition is challenging due to the rapid transition between the events, occlusions and fast movements of the players, varied camera viewpoints, and camera motions. Moreover, spatiotemporal structures of events vary greatly in different sports. For example, the locations of the players in volleyball is relatively static compared to players in ice hockey. All these factors make the recognition problem complex and hard to generalize across different sports. Previous work have tried to address these issues to some extent but most group activity classification problems have targeted non-sports applications [2, 7, 19, 6]. Consequently, studying a particular sport with domain knowledge is valuable and complementary to general activity recognition.

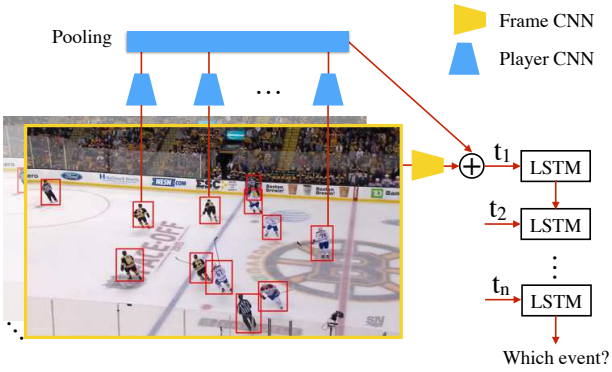

Figure 1. Our method can classify puck possession events for ice hockey games from a sequence of images. We first extract features from a whole frame and individual players. Then we train an LSTM model using a sequence of temporal features.

We are interested in puck possession events classification in ice hockey games. The classification results may enable coaches and analysts to determine both strategic concepts and support evaluation of individual players. Unlike other ball sports such as basketball and soccer, the puck in hockey can be in the possession of neither team for extended time periods, for example, when the puck moves out from the defensive zone or into the offensive zone ('dump out' and 'dump in').

The playing surface (rink) is large and enclosed by the boards. Play-by-play commentary, provided in real-time during a broadcast, annotates shooting, scoring and hit events at time intervals of 3 to 15 seconds, typically, while the intervals between our annotated events range from 3 to 200 frames (30 FPS). The output of our system can greatly benefit the manual events annotation with some tolerance of error.

Our proposed model aims to classify five puck posses-

(a) Dump in (b) Dump out (c) LPR

(d) Pass (e) Shot

Figure 2. Example images of puck-possession events. In each image, the dashed red line is the potential trajectory of the puck except for LPR in which the red lines are the trajectories of potential player movements.

sion events which are *dump in*, *dump out*, *pass*, *shot* and *loose puck recovery* (LPR). The descriptions of these events are in Table 1. Figure 2 shows example images of these events. In the images, the dashed red lines show the movements of pucks or players. They also show the importance of temporal information in events classification. Figure 3 shows the schematics of these events on the ice hockey strategy board.

Our model uses a deep architecture which only requires detected bounding boxes of the players and corresponding frames during test, simplifying the input of the system. Figure 1 shows the pipeline of our method. The input of our method is a sequence of images with player detection results (bounding boxes in the image). Our method first extracts context features from a whole image, and individual features from player image patches. Because the number of detected players changes over frames, we use a max pooling layer to aggregate the individual features. Then, we use an LSTM model to train an event classification model using features from the sequence of given images.

The main contribution of our work is two-fold. First, we propose a benchmark for event classification on a new challenging ice hockey dataset. Second, we extensively study the features from whole frames, individual players and temporal information. We provide solid evidence that our model works best when individual's information is combined with the context of the events in ice hockey games.

2. Related work

Group activity classification Group activity classification has been an active area of research over the years. Most previous work has relied on hand-crafted features. For example, Amer *et al*. [2] proposed a Hierarchical Random Field (HiRF) model that captures long range temporal dependencies between latent sub-activities and infers the activity class at the root node of the graphical model. Choi and Savarese [7] proposed a unified framework to track multiple individuals, identify individual actions, understand the interactions and identify the collective action. Lan *et al*. [18] proposed a hierarchical model that represents low-level actions, social roles and scene-level events. Lan *et al*. [19] proposed a latent variable framework to capture individual actions, interactions among the individuals and group activities. Ramanathan *et al*. [24] focused on identifying social role of individuals in a weakly supervised approach where the inter-role interactions are modeled using a Conditional Random Field (CRF). Recently, Shu *et al*. [27] uses a spatiotemporal AND-OR graph to jointly infer groups, events and human roles in aerial videos.

Deep learning and sports video analysis As more data became available, the success of Convolutional Neural Networks (CNN) has been proved in numerous applications over the last decade on computer vision tasks such as image recognition [17] and video analysis [15, 28]. Recurrent Neural Networks (RNN) particularly Long Short-term Memory (LSTM) [12] are widely popular models that are well suited for variable length sequence inputs. LSTM has been successfully applied to speech and handwriting recognition [10], human action recognition [9, 3] and image caption generation [31, 14].

On the other hand, the domain of sports analysis is vast because each sport has its unique characteristics. As a result, a model from a particular sport is unable to achieve reasonable performance on other sports dataset. However,

researchers have narrowed down the domain into specific sports particularly the most challenging ones and have tried to solve different aspects over the years. For example, Yue-Hei Ng *et al*. [33] proposed and evaluated different deep neural network architectures to classify longer sequence of sports videos. Meanwhile, various approaches have been proposed for different types of sports analysis [4, 21, 26].

Researchers have shown different ways to combine LSTMs with CNNs or graphical models for group activity recognition. For example, Deng *et al*. [8] integrated a graphical model into a deep neural network. The network learns structural relations by representing individuals and the scene as nodes passing messages among them and imposing a gating mechanism to determine the meaningful edges. But the method is not designed for sports activities where interactions between individuals are generally more complicated. Ibrahim *et al*. [13] built a hierarchical deep network to learn the individual actions using one LSTM which is then combined with features extracted from CNNs to pass into another LSTM to predict the group actions. However, their method has difficulties on events which are very similar to each other. Moreover, their model needs explicit labels for individual actions which are expensive and hard to label for sports like ice hockey. In [23], Ramanathan *et al*. argued that in many group activities redundant information can be ignored by concentrating on a subset of people who contribute to the group activity. Thus, they first extracted features from individuals who are 'attending' to the event as well as global context features representing the entire scene and then solved the problem of event classification using a deep network.

Building upon the existing work, our model takes advantage of the discriminative power of deep learning and captures the structural and spatiotemporal information in group activities. Moreover, it shows how combining contextual information and person level features can improve accuracy for events that are very similar to each other.

3. Our method

Visual cues The most descriptive cue for possession events classification is the location of the puck and the location of players in playing ground coordinates. However, it is extremely hard to track the puck in images as the puck is very small and moves very fast. Due to motion blur, the puck's color and texture can be merged into backgrounds. Without the puck information,the event classification problem is even more difficult. Estimating player locations in the playing ground coordinates requires the camera parameters, which is also challenging for fast moving cameras.

Alternative cues for puck possession events classification are player locations in the image and their spatiotemporal information. Players are coached to keep the team shape and move to offense/defense together. But looking only at individual players can be ambiguous in some events. For example, the player appearance and action might be very similar in the two events 'pass' and 'shot'. In this case, additional cues such as context information are necessary to distinguish these two events.

Method overview The input to our model is a sequence of images as well as the player bounding boxes in each image. The bounding boxes of players can be in any order as our method does not require player trajectories. The output of the model is a group activity label for the entire sequence.

Our method has two parts: feature representation and events prediction. In the feature representation, we aggregate different types of features that are extracted from a pre-trained convolutional network. In the events prediction, we use a single layer LSTM model. Our main efforts are on integrating different types of features to improve classification accuracy.

3.1. Individual and context information

We use appearance features to model individual players. The appearance feature is extracted by the $fc7$ layer of AlexNet [17] using the sub-image of a player. We choose to use the pre-trained AlexNet (on ImageNet object recognition task) because it has been successfully used in various computer vision tasks [9, 13] and we only have a small number of training/testing data. The output of the CNN represents the appearance information of an individual player.

Interaction among the individuals is essential to determine group activities. However, it is difficult to incorporate the exact location of individuals into deep features as a cue, which requires player locations in playing ground coordinates. To solve this problem, we use max pooling the features of individual players in a particular frame to incorporate player interactions.

We use deep features from a whole image to model the context information. In each frame, we use the $fc7$ activation in AlexNet as the representation of the context. The intuition for adding this context is that some events can only be determined if we know the scene information. For example, if we consider the events 'dump in' and 'dump out' in ice hockey, they are almost the same except the fact that they occur at different zones.

3.2. Temporal model

We use a Long Short Term Memory (LSTM) network [12] to model the temporal information of the images. In every timestep t, the LSTM includes a hidden unit h_t, an input gate i_t, forget gate f_t, output gate o_t, input modulation gate g_t and memory cell c_t. The LSTM formulation can be represented as the following equations:

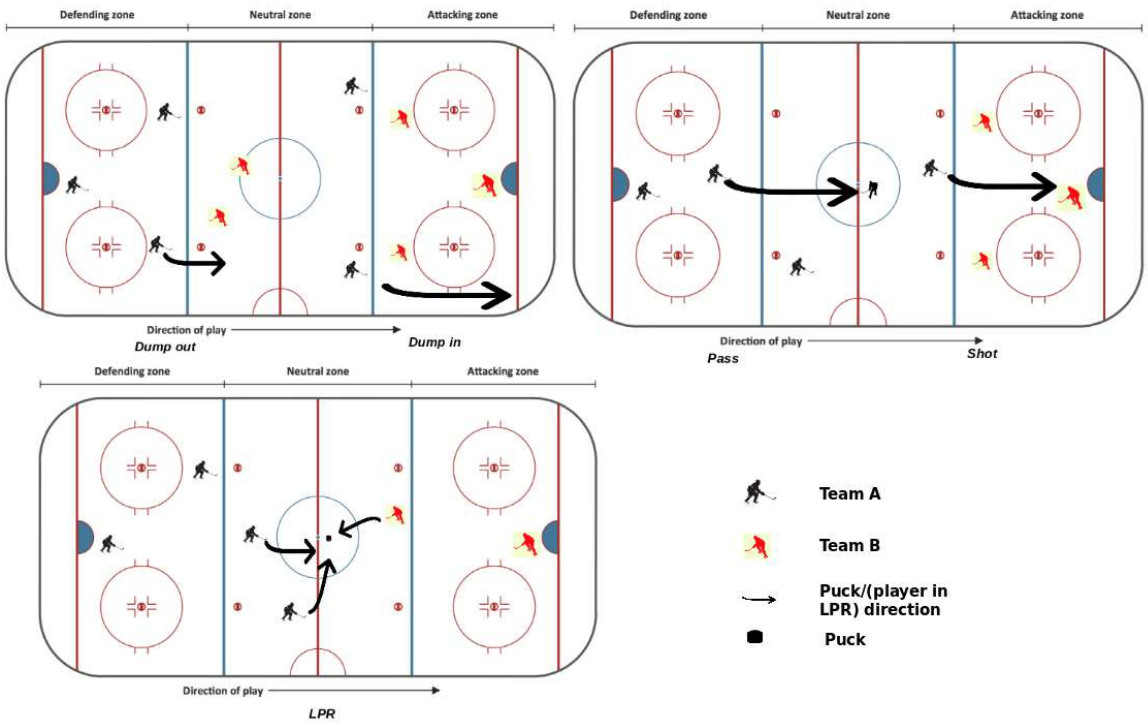

Figure 3. Schematics of puck possession events. This figure shows schematics of five puck possession events that our system aims to classify. In some events, individual players appearances/motions could be very similar such as 'dump in' and 'dump out'.

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
g_t &= \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \phi(c_t)
\end{aligned} \tag{1}
$$

where W terms denote weight matrices (*e.g.* $W_x i$ is the matrix of weights from the input to the input gate), b terms are bias vectors. σ is the logistic sigmoid function, ϕ is the $tanh$ function, $\odot$ is the element-wise product. We have also tried more advanced LSTMs such as the LSTM with peephole connections [25]. There is no performance gain compared with the basic LSTM model.

In this representation, the group dynamics is evolving over time and the event at a given time instance can be determined based on the hidden state computation from the preceding states and current input x_t.

4. Experiments

We conducted experiments on an ice hockey dataset. In this dataset, an example contains a target frame which is associated with an event label. The target frame is generally a frame that marks the beginning of an event. The length of the sequence is varying from 2 to 24 frames. For the feature extraction, we use the AlextNet pre-trained from ImageNet for object detection if not specified otherwise. For the prediction model, we use SoftMax classification and LSTMs.

4.1. Baselines

We considered the following baseline models for the evaluation:

1. Frame-level Classification with CNN (M1): This baseline extracts frame level features from target frames and classifies the event for the target frames using Softmax.

2. Person-level Classification with CNN (M2): This baseline first extracts player level features from target frames. Then it max pools across players and then classifies event for target frames using Softmax.

3. Frame-level Temporal Model 1 (M3): This is an extension of the first baseline (M1). Instead of using the target frames and Softmax classification, this method feeds the frame level features from the whole sequence into an LSTM to classify events for the whole sequence.

4. Person-level Temporal Model (M4): This is an extension of the second baseline (M2). It feeds the player

Event	Description	# Examples
Loose puck recovery (LPR)	The player recovered the puck as it was out of possession of any player	1,071
Pass	The player attempts a pass to a teammate	1,057
Shot	A player shoots on goal	145
Dump in	When a player sends the puck into the offensive zone	95
Dump out	When a defending player dumps the puck up the boards without targeting a teammate for a pass	139

Table 1. Event descriptions and corresponding number of training examples in the ice hockey dataset.

Method	Fine-tuning	Accuracy (%)
M1	w/o	31.7
M2	w/o	39.7
M3	w/o	39.6
M4	w/o	42.4
M5	w	42.1
M6	w	46.8
C3D [30]	w	44.0
Our method	w/o	**49.2**

Table 2. Performance of our model on Ice Hockey compared to the baselines. In the fine-tuning column, w/o and w represent without and with fine-tuning, respectively.

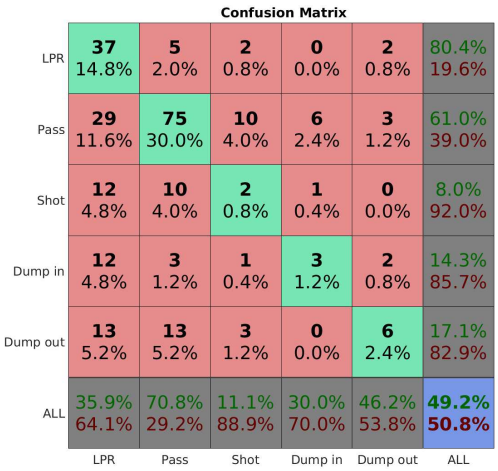

Figure 4. Confusion matrix of event prediction of our method on the ice hockey dataset.

level features from a sequence of images to an LSTM to classify events for the whole sequence.

5. Frame-level Classification with fine-tuned CNN (M5): This baseline is similar to M1 but we fine-tuned the AlexNet using the target frame events.

6. Frame-level Temporal Model 2 (M6): This baseline is same as M3 but it uses fine-tuned Alexnet features rather than pre-trained features.

We also compare our method with the C3D [30] network. The C3D network is pre-trained on the UCF101 action recognition dataset [29] then fine-tuned in our dataset.

4.2. Ice hockey dataset

This dataset consists of National Hockey League(NHL) videos and was obtained from SportLogiq. We used part of this dataset and considered five puck possession events. Table 1 shows the descriptions of the events and corresponding number of instances in the dataset. It clearly shows that some of the events occur very rarely such as 'dump in'. We randomly used $2,507$ events for training and 250 events for testing. The dataset has the annotated frame numbers where an event occurred and we used the preceding frames for our temporal classification. All the events are considered to be independent of each other and were trained as individual short clips of events, which is a standard protocol used for activity recognition tasks [13][30] .

4.3. Quantitative Results

Table 2 shows the classification accuracy of our model and the seven baselines. Our method outperforms all the baselines by a distinct margin. Among the baselines, the player level features work better than the frame-level features when we use pre-trained model and the LSTM model performs better than the SoftMax. For example, M2 performs slightly better than M1 because some of the background features in M1 may be uninformative in regions such as the unused portion of the rink or the crowd. Moreover, the AlexNet is pre-trained from image recognition task in which sample images usually have an object in the middle of the frame. When we apply this pre-trained model to our problem, it is more suitable for extracting player level features as players are in the middle of the frame. On the other hand, a whole frame has multiple players and large areas of background, thus the pre-trained AlexNet becomes less suitable. That is why we have fine-tuned AlexNet for frame-level classification using the target frame events. M5 and M6 show that the fine-tuned frame features outperform

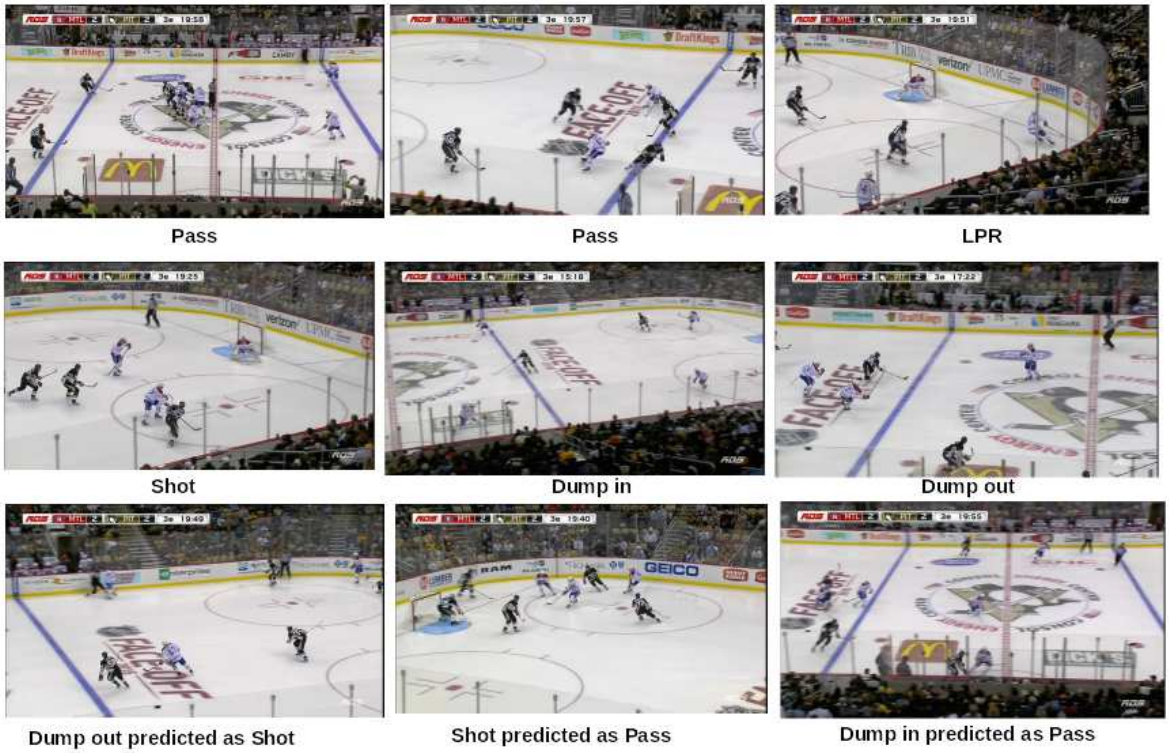

Figure 5. Qualitative results. The first and second rows show the examples that are correctly recognized by our model. The bottom row shows the mistakenly predicted examples.

both M1 and M3. However, we had challenges to fine-tune our final model because we do not have person-level annotations that could be used to fine-tune player level features. Our method is also better than the C3D network, like [23] for activity recognition in basketball games.

The results also indicate that capturing spatial features of individual players is necessary as they represent different actions of the players. For example, a player having the puck acts differently from a player who is far away from it. It is also seen that adding temporal information to both M1 and M2 improves the performance. This indicates that the spatial features evolve over the time as they approach the target events. All the baseline models show that adding either the frame level features or the player level features do quite well in predicting the events. Adding frame level features is important because events such as 'dump in' and 'dump out' can possibly be distinguished only by zone information.

Figure 4 shows the confusion matrix of the events. The model performs well on 'dump out' and 'pass'. On the other hand, the model has lower accuracy on 'dump in' and 'shot'. Moreover in many cases it confuses 'LPR' with the 'pass' event. One explanation is that players appearances are similar in the image for these three events and our method does not encode player location information in the feature. Overall, the less frequent events perform poorly compared to the dominant events because the training dataset is highly imbalanced. One way to fix this problem is to collect more training instances for these events.

4.4. Qualitative results

Figure 5 shows qualitative results of event prediction. Our method successfully recognized some challenging examples of 'pass' and 'LPR' events. However, our method had difficulties in some cases. For example, it has failed to distinguish some 'dump in' and 'shot' events from 'pass' events. The reason is 'LPR' and 'pass' are the most frequent events whereas 'dumps' and 'shots' occur very rarely.

4.5. Implementation details

We extract deep features in Matlab using an AlextNet pre-trained on the ImageNet for object recognition task. The classification models are trained using the Tensorflow [1] framework. Our LSTM network consists of 1,000 hidden nodes, 500 input features and optimizes softmax cross entropy loss function. We used decreasing learning rate, dropout for regularization, batch-normalization and Adam optimizer [16].

5. Conclusion

In this paper, we proposed a deep learning model to classify group activity in ice hockey. We have shown that feature aggregation is essential in determining events. The five puck possession events can be classified without the need of explicit labeling for individual actions or puck information. Future work will focus on incorporating player motions, an attention mechanism and 3D pose information into this model. Another important extension would be finding motion and position of hockey sticks with respect to the players as well as taking advantage of gaze information of the players.

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Institute for Computing, Information and Cognitive Systems (ICICS) at UBC, and enabled in part by WestGrid and Compute Canada

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 6

[2] M. R. Amer, P. Lei, and S. Todorovic. HIRF: Hierarchical random field for collective activity recognition in videos. In *European Conference on Computer Vision (ECCV)*, 2014. 1, 2

[3] F. Caba Heilbron, W. Barrios, V. Escorcia, and B. Ghanem. SCC: Semantic context cascade for efficient action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[4] D. Cervone, A. DAmour, L. Bornn, and K. Goldsberry. POINTWISE: Predicting points and valuing decisions in real time with nba optical tracking data. In *8th Annual MIT Sloan Sports Analytics Conference*, 2014. 3

[5] J. Chen, H. M. Le, P. Carr, Y. Yue, and J. J. Little. Learning online smooth predictors for realtime camera planning using recurrent decision trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. *European Conference on Computer Vision (ECCV)*, 2012. 1

[7] W. Choi and S. Savarese. Understanding collective activities of people from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1242–1257, 2014. 1, 2

[8] Z. Deng, A. Vahdat, H. Hu, and G. Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 3

[10] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2014. 2

[11] A. Hilton, J.-Y. Guillemaut, J. Kilner, O. Grau, and G. Thomas. Free-viewpoint video for TV sport production. In *Image and Geometry Processing for 3-D Cinematography*, pages 77–106. 2010. 1

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 3

[13] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 5

[14] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[16] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 2, 3

[18] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1361, 2012. 2

[19] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, 2012. 1, 2

[20] J. Liu, P. Carr, R. T. Collins, and Y. Liu. Tracking sports players with context-conditioned motion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[21] B. Macdonald. An improved adjusted plus-minus statistic for NHL players. In *Proceedings of the MIT Sloan Sports Analytics Conference*, 2011. 3

[22] T. B. Moeslund, G. Thomas, and A. Hilton. *Computer vision in sports*. 2015. 1

[23] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei. Detecting events and key actors in multi-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 6

[24] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[25] H. Sak, A. W. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *In Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014. 4

[26] O. Schulte, M. Khademi, S. Gholami, Z. Zhao, M. Javan, and P. Desaulniers. A markov game model for valuing actions, locations, and team performance in ice hockey. *Data Mining and Knowledge Discovery*, pages 1–23, 2017. 3

[27] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[28] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, 2014. 2

[29] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 5

[31] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[32] X. Wei, P. Lucey, S. Morgan, and S. Sridharan. Forecasting the next shot location in tennis using fine-grained spatiotemporal tracking data. *IEEE Transactions on Knowledge and Data Engineering*, 28(11):2988–2997, 2016. 1

[33] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3