Interaction Classification with Key Actor Detection in Multi-Person Sports Videos

Farzaneh Askari[1] Rohit Ramaprasad[2] James J. Clark[1] Martin D. Levine[1]

[1]McGill University, Montreal, QC, Canada
[2]Birla Institute of Technology and Science, Pilani, Rajasthan, India
{farzaneh.askari,james.clark1,martin.levine}@mail.mcgill.ca, f20180224@pilani.bits-pilani.ac.in

Abstract

Interaction recognition from multi-person videos is a challenging yet essential task in computer vision. Often the videos depict actions with multiple actors involved, some of whom participate in the main event, and the rest are present in the scene without being part of the actual event. This paper proposes a model to tackle the problem of interaction recognition from multi-person videos. Our model consists of a Recurrent Neural Network (RNN) equipped with a time-varying attention mechanism. It receives scene features and localized actors features to predict the interaction class. Additionally, the attention model identifies the people responsible for the main event. We chose penalty classification from ice hockey broadcast videos as our application. These videos are multi-persons and depict complex interactions between players in a non-laboratory recording setup. We evaluate our model on a new dataset of ice hockey penalty videos and report 93.93% classification accuracy. We include a qualitative analysis of the attention mechanism by visualizing the attention weights. Our code is publicly available [1].

1. Introduction

Human Action Recognition from Videos (HARV) has drawn a significant amount of attention to the field of computer vision due to its diverse applications, such as video surveillance, health care, entertainment, and sports analytics. It owes its popularity to the increased usage of cameras and communication and sharing platforms such as social media and broadcasting. Additionally, the development of powerful computing technologies such as Graphics Processing Units (GPUs) along with the emergence of deep learning architectures marked a great leap forward in the evolution of HARV research. HARV is an inevitable component of video analysis due to the fact that human actions and their interactions with the environment account for the predominant part of interactions in the videos.

Figure 1. Sports broadcast scenes are often crowded. In this figure, the two players in the bounding box are the "key" actors responsible for the main interaction (penalty). The goal is to classify the main event while attending to the key actors.

A significant portion of HARV literature focuses on single-person actions, where an individual performs a primary action. However, in many applications, the activity involves the interaction between two or more individuals. The majority of datasets [33, 35, 55] are centered around simple interactions (e.g., handshake) recorded in a controlled laboratory environment. Often the cameras are adjusted to provide a distinguishable view of the action (and actor), minimize the occlusions, and remove the camera motion. Addressing the simple scenarios resulted in progress in HARV over the years; however, nowadays, a vast majority of HARV applications demand automated systems capable of recognizing complex actions in naturally occurring environments.

In multi-person videos, the scene contains several actors, with only a few of them involved in the primary event. For example, in a penalty scene from a sports broadcast video,

[1]https://github.com/SummerVideoAnalysis/Interaction-Classification-with-Key-Actor-Detection-in-Videos

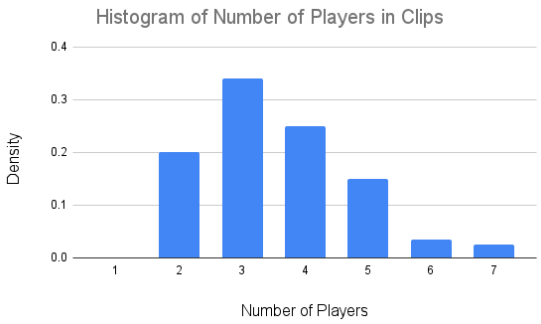


Figure 2. Histogram of maximum number of players in each clip across the ice hockey penalty dataset

(usually) only two players are involved in the penalty, while other players present in the frame are not taking part in the key interaction Fig. 1. Following [29] in this context, we refer to the actors responsible for the primary event, the "key" actors. Therefore, in multi-person videos, it is crucial to identify the key actors as well as recognize the primary interaction. Our proposed methodology in this paper tackles the problem of key actor detection from multi-person videos without requiring explicit annotations of the key actors.

Sport broadcast videos offer a variety of challenges in terms of multi-person HARV. In recent years the research community has contributed to the two main categories of this field. The mainstream contribution is to recognize and distinguish different sports activities (e.g., swimming versus running) using sports datasets such as Sports-1M [24] or UCF Sports [26]. Most of the methods in this branch leverage the inter-class appearance (context, e.g., swimming pool vs. running track) differences to achieve good performance. The second category of sports analysis recognizes the players' activities or the team during a game, which is more challenging and less studied. The available studies in this field are often sparse and sport-specific because each sport has unique spatiotemporal dynamics and characteristics. This domain of sports analysis from broadcast videos provides beneficial insight into the game for coaches, referees, and sports analysts. Features such as players' appearances (including pose), players' positions and their trajectories, and the trajectory of sport-specific objects (e.g., the puck in ice hockey) are prevalent in this domain.

Among the sports, ice hockey (hockey for short) broadcast videos include complex scenes and interactions due to the frequent (self) occlusions, fast movements of the players in a relatively small rink (i.e., $\frac{1}{4}$ of a soccer field), varied camera viewpoints (i.e., scale and angle), blurry scenes due to camera motion, and rapid transitions between game events. Additionally, the bulkiness and the color of the players' uniforms create confusion while extracting the appearance features such as the human skeleton pose (pose for short). For example, in the far camera shots, the coloring of the jerseys can make the players blend in with the background.

Penalties are examples of complicated human interactions during a sports game that can significantly affect the dynamics and directions of the game. The players' speed and density on the hockey rink cause much physical contact, making penalties an inevitable part of hockey games. The substantial (self) occlusions in penalty scenes, the sub-optimal and varying camera viewpoints (i.e., scale and angle), along with a low inter-class variance of penalty scenes, call for a novel framework to address these challenges. Therefore, this paper proposes a CNN-RNN based model equipped with an attention mechanism that recognizes penalties from ice hockey broadcast videos while isolating the responsible players for the event.

Our model receives an ice hockey penalty clip, players' pose, and hockey stick annotation (i.e., coordinates of stick ends) as input and outputs a penalty class. The video frame CNN features along with pose information is input to multiple RNNs (LSTMS) dedicated to frame features, pose features, and event. A time-varying attention mechanism is employed to identify the important (key) players based on the information from previous time steps. The local players' features along with the global video features are used for penalty classification. Additionally, the key actors are implicitly detected within the attention mechanism.

Our paper is structured as follows. Sec. 2 reviews the current literature related to our work. Sec. 3 presents our dataset and the accompanying annotations. We elaborate on different components of our methodology in detail in Sec. 4. Sec. 5 discusses our experimental setup, classification results, and analysis of attention mechanism. Finally, we conclude the paper in Sec. 6.

Figure 3. An example of pose estimation failure to capture accurate pose on our ice hockey penalty data.

2. Related Work

The emergence of deep learning marks an era of advancement in many computer vision applications, including HARV. Simonyan et al. pioneered a popular branch of HARV by proposing two-stream architecture for video classification [37]. The idea is based on the fusion of two CNNs, with one CNN learning representation from the stack of RGB images (i.e., video frames); and the other capturing motion information from image-like modalities (e.g., optical flow). The success of this proposition inspired many researchers to explore variations of CNN [6, 21, 43, 51] and RNN [10, 11, 38, 54, 57] architectures for the task of HARV.

The pose is a compact feature that summarizes an image (or a video frame) into important key-points, the coordinates values of human joints. Many successful methods for HARV integrate pose as one of their features in combination with RNN [25, 31, 44] and CNN [8, 9, 16] architectures or combine it with appearance and motion features [58]. Although the pose is a useful feature for HARV, it is expensive and time-consuming to annotate. The advancement of successful pose estimators such as OpenPose [5] alleviated the need for manual annotation of pose and facilitated its integration into HARV.

Interaction recognition has been recently the subject of a handful of studies. Some studies propose an RNN architecture that takes CNN extracted features as input to perform temporal reasoning of the individual interactions [10, 25, 34, 36]. Similar to HARV, interaction recognition studies integrated pose features in their model, either explicitly as the main feature [25, 31, 44] or as a guide to extract Spatio-temporal features [22, 23, 52, 56].

Person images and videos are found under various names such as key actor detection [7, 29], important people detection [19], and action localization [39]. It is important to note that even though many studies address spatial action localization from videos, their target dataset includes actions defined around a single person [17, 20, 30]. In terms of supervision, some studies rely on the expensive and time-consuming annotation of the key actors [7, 27, 28, 41, 50]. Among the key actor detection models from multi-person videos, the study by Ramanathan et al. [29] is closer to our approach. They address the problem of key actor detection in basketball broadcast videos using only weak supervision. The paper uses frame-level CNN representations and the spatially localized frame-player-level features to input an RNN model with an attention mechanism.

The literature on Ice hockey is not extensive. Many studies focuses on player identification and tracking [45, 48, 49]. Another stream of studies focus on localizing spaces [14] and objects (e.g., puck) [46, 47] from the videos. Neher et al. proposed a CNN model for the pose estimation of hockey players. Studies such as [4, 13] tackle the problem of recognizing single-person actions (e.g., passing, shooting) using pose and optical flow. Tora et al. [42] propose a CNN-RNN based method to classify multi-person puck possession events (e.g., dump in, shot). To our knowledge, there is no study available on multi-person penalty classification from ice hockey broadcast videos.

3. Dataset

Our ice hockey penalty dataset consists of three Slashing, Tripping, and No Penalty classes with 76, 80, and 98 clips, respectively. The clips are collected from National Hockey League (NHL) broadcast videos [3]. We select Slashing and Tripping because they are among the top occurring penalties during hockey games [2]. We add a No penalty class that includes snippets of players skating, face-off, goal, and other game events except penalties.

To label the two penalty classes, we utilize the play-by-play tool provided by NHL. NHL [3] defines slashing and tripping penalties as follows: "Slashing is the act of a player swinging his stick at an opponent, whether contact is made or not" and " a player shall not place the stick, knee, foot, arm, hand or elbow in such a manner that causes his opponent to trip or fall ".

The clips are two to six seconds long and include two to seven players. The clips from penalty classes have exactly two key actors. The total number of players in each clip can vary from frame to frame, meaning that all the players (key or others) can exit/enter the scene anytime during the event. Fig. 2 demonstrates the histogram of the maximum number of players across the dataset.

The dataset represents challenges such as significant view variation of the penalty scenes in terms of angle and scale (i.e., close to medium-far shots), camera motion, blurry frames, and (self) occlusion. The penalties are presented either in actual speed or slow-motion replays.

We include ground-truth pose annotation of 14 key-points for all the clips. We first take advantage of pose extraction libraries (i.e., Openpose [5]). Often the pose extractors are trained on "usual" human poses such as standing and sitting. Consequently, a complex interaction (e.g., penalties) that includes (self) occlusions and "unusual" poses is challenging for the pose extractor, which in turn leads to numerous missing joints in the frames. Fig. 3 demonstrates an example of pose estimation failure to capture precise pose on our dataset.

Therefore, we post-process the extracted poses and fill in the missing joints using manual annotation and heuristic conditions based on the pose information from the neighboring frames. In our annotations, if the player (or referee) is too far (i.e., more than half the shot length) or mostly invisible (i.e., more than half of the joints out of frame), we exclude their poses. Additionally, for the joints that are out of frame, we set their coordinates to zero. In our dataset, we avoid redundant identification number assignments. It

Figure 4. Frames from our ice hockey penalty dataset, including pose and hockey stick annotations. Classes from top to bottom: tripping, slashing, no penalty.

means, that if players leave and re-enter the scene, their previous ID will be assigned to them; and while they are absent from the scene their joint coordinates are set to zero.

Many penalties include the use of a hockey stick. This information is crucial to classify penalty events. Therefore, our dataset contains ground truth annotations of hockey grip and heel for every player in each clip. We use CVAT annotation tool [1] for this purpose. Fig. 4 shows a few examples of clips and their pose annotations in our dataset.

4. Method

We propose a CNN-RNN based model that receives video frames and poses as input and classifies each video into three classes. The model is equipped with a time-varying attention mechanism on the players' pose. This section elaborates on feature extraction, network achitecture, and attention mechanism. Additionally, we introduce two variations of our primary model as baselines that only receive frames and pose annotations as input, respectively. Fig. 5 demonstrates an overview of our proposed model. The notations in this section are inspired by the work of Ramanathan et al. [29].

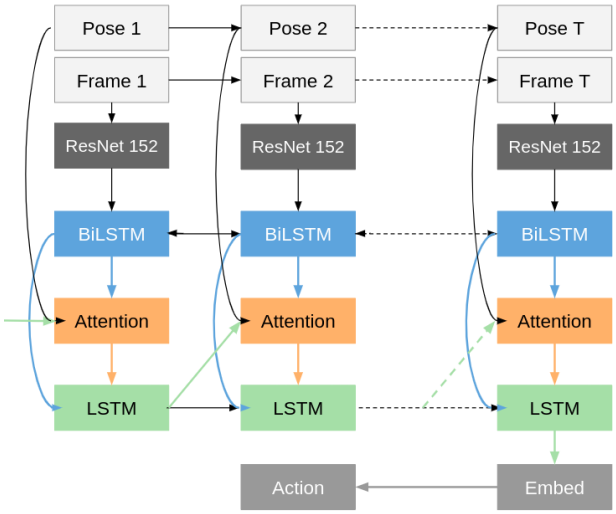


Figure 5. The architecture of our model. The scene features are extracted from each video frame. A BiLSTM extracts global information from the scene features. These features and players' poses are input to an attention model. The attention mechanism's output and scene features are input to the interaction classification LSTM for final interaction classification.

4.1. Feature extraction

In order to extract scene features from video frames, we feed every frame to a Resent152 network [18]. The features from the last convolutional layer of Resent152 are then embedded into a 512-dimensional vector. We call this vector f_t. Additionally, we use the players' pose and hockey stick annotation as localized features for every frame. p_{ti} is a 32 dimensional vector representing x, y coordinates of pose and hockey stick key-points (16 in total) for player i at time-

step t.

4.2. Network and attention mechanism

We extract the global context feature from f_t using a bidirectional LSTM, notated as BiLSTM$_f$. At every time frame, the hidden state from BiLSTM$_f$ is input to an attention mechanism and is part of the input to an LSTM, which classifies the interaction, LSMT$_c$. The second part of the input to LSMT$_c$ is the output of our attention mechanism that applies attention to players' features in each frame, based on the previous hidden state of LSMT$_c$ (h^c_{t-1}) and the current hidden state of BiLSTM$_f$ (h^f_t).

$$h^f_t = \textbf{BiLSTM}_f(h^f_{t-1}, h^f_{t+1}, f_t) \tag{1}$$

$$h^c_t = \textbf{LSTM}_c(h^c_{t-1}, h^f_t, pa_t) \tag{2}$$

The inputs to the attention model at time-step t are h^c_{t-1}, h^f_t, and players features p_{ti} where $i = 1, ..., N$ with N representing the maximum number of players in the video. First, each player's feature (p_{ti}) is concatenated with h^c_{t-1} and h^f_t and then fed to a Multi-Layer Perceptron (MLP). The output of the MLP is masked and Softmaxed to generate attention weights for each player's feature. The final representation from the attention mechanism is the weighted sum of players' key point features present in frame t. We denote this vector by pa_t.

$$pa_t = \sum_{i=1}^{N_t} \textbf{Softmax}\left(MLP([p_{ti}, h^f_t, h^c_{t-1}])\right) p_{ti} \tag{3}$$

Finally, as mentioned earlier, the concatenation of pa_t and h^f_t are input to the interaction classification LSTM. The output of LSMT$_c$ passes through an embedding layer, and a Cross-Entropy loss is calculated on the predictions from the last time-step (T) of the LSMT$_c$ (noted as $y^c_{T_i}$) . In equation 4, K represents the number of classes. In our ice hockey dataset $K = 3$ (i.e., tripping, slashing, and no penalty) Fig. 5 demonstrates our architecture.

$$loss = -\sum_{k=1}^{K} y^c_{T_i} \log y^c_{T_i} \tag{4}$$

4.3. Baseline models

As baselines, we introduce two simpler versions of our model. Each baseline utilizes only one of the available features (i.e., global frame features or localized players features). Our first baseline model (Model 1) receives the scene features as input to the BiLSTM$_f$ followed by the LSMT$_c$ to output labels. Therefore, this model does not have access to pose information and cannot localize the key actors.

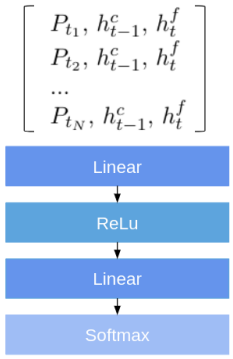


Figure 6. Attention model. At time-step t, the attention model identifies the key players based on h^c_{t-1} and h^f_t.

On the other hand, our second baseline model (Model 2) takes in the players' poses. The poses and the hidden states from LSMT$_c$ are input to the attention mechanism. The weighted sum of players' poses is the output of the attention model and input to the interaction classification LSTM. It means, in our second baseline, neither the attention model nor the interaction LSMT$_c$ has access to scene features from BiLSTM$_f$.

5. Experimental Evaluation

5.1. Experimental setup

The hidden states dimensions in LSTM and BiLSTM are 512. The MLP of the attention mechanism includes a linear layer of dimension 512, ReLu activation, linear layer of dimension 1, followed by a Softmax. We sample 64 frames from the clips. The batch size is 32, with a learning rate of 0.00005 for the primary model (Model3). Additionally, we augment our dataset using the horizontal flip and affine transformation (e.g., scale). We use 70%, 20%, and 10% of our data for training, validation, and testing and choose the hyper-parameters based on the loss and metric on the validation set. We run each model using 3 different initial seeds and report the test accuracy averaged over the seeds.

5.2. Results

In this section, we present the results of our fusion model (Model3) and compare them against our baseline models (Model1 and Model2). We report the accuracy of the (penalty) event classification task; meaning, the number of times that our model correctly predicts the label (i.e., tripping, slashing, no penalty) of the input video. The event classification accuracy is shown in Table 1. Model3 that fuses the scene and localized players' features outperforms both of the baselines models with access to only one of the modalities, emphasizing the importance of integrating global and local information.

Model	Accuracy (%)
Model1: only frames (no Att)	87.43
Model2: only pose (Att)	80.66
Model3: frames and pose fusion (Att)	**93.93**

Table 1. Penalty classification accuracy for our interaction classification primary and baseline methods

Model	Accuracy (%)
Model2: only pose (Att)	**80.66**
Model2 wo stick: only pose (Att)	74.86
Model3: frames and pose fusion (Att)	**93.93**
Model3 wo stick: frames and pose fusion (Att)	90.46

Table 2. Studying the effect of stick keypoints on penalty classification

Many penalties in ice hockey involve using the hockey stick; therefore, the knowledge of the location of the hockey stick plays an essential role in the classification of penalties. In Table 2 we show an ablation study on the effectiveness of hockey stick annotations. Specifically, we run Model2 and Model3 with the same settings, except the number of key-points; we feed only the players' body key-points and exclude the hockey stick key-points (i.e., feeding 14 key-points in total). Table 2 demonstrates the performance drop upon excluding the stick key-points. In Table 3 we compare our method against a few popular action/interaction recognition approaches. It is important to note that, even though these methods classify the interactions, they do not localize the key players. However, our proposed method offers interaction classification as well as key actor detection. Long-term Recurrent Convolutional Networks (LRCN) [10] is a CNN-RNN based approach that receives video frames as input. PoseC3D [12] is a 3D-CNN-based model that creates 3D heatmap stack representations from human skeleton data as input. The pose estimation is an element of their approach; therefore, we use their proposed pose estimation to acquire the raw human skeleton. To do so, we first use Faster-RCNN [32] to detect actors, followed by HR-Net pose estimation [40], which is a top-down pose estimator. To train the PoseC3D model, from available back-ends, we pick Slowonly [15] with Resnet [18] backbone. Finally, we run Spatial temporal graph convolutional networks (ST-GCN) [53] as a member of popular graph convolutional networks.

5.3. Analyzing attention

In Fig. 7, we visualize the performance of the attention mechanism in terms of detecting the key actors and their relation to interaction prediction. We extract the attention weights on the test set and round up the number, meaning minimal attention weights are rounded to zero and not displayed. The weights have passed through a masking mechanism as well as Softmax. (i.e., Eq. (3)).

Model	Accuracy (%)
LRCN [10]	63.64
ST-GCN [53]	67.35
PoseC3D [12]	81.63
Ours	**93.93**

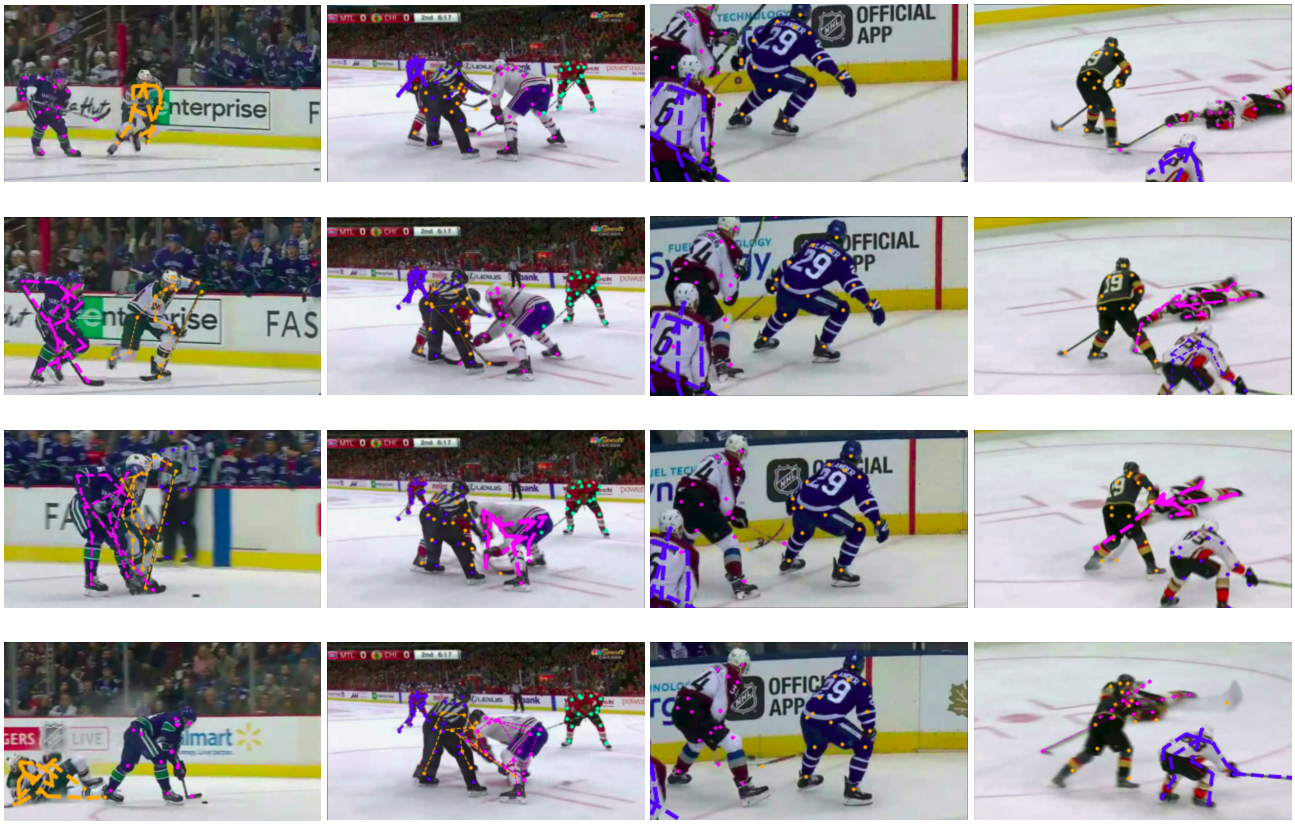
Table 3. Comparing our primary model with some popular action recognition methods. Note: none of the methods above, offer key actor detection

In Fig. 7 (a), the attention model correctly attends to two key players while ignoring the non-key actor (i.e., the referee). Toward the end of the clip, the model places heavier weight on the falling player, which is a good indicator of a penalty interaction. As depicted in Fig. 7 (b), Our model correctly recognizes a No penalty interaction by exploring different players present in the scene over time and focusing on the two players ready for a face-off. However, in the cases of Fig. 7 (c), the attention is fixated on the wrong player when the main event is finished resulting in incorrect classification. Fig. 7 (d), depicts an unusual case of the player committing the penalty falling on the rink. Additionally, the model attends to the non-key actor during the penalty peak resulting in an incorrect classification.

Overall we can observe the relation between key actor localization and the interaction classification, demonstrating the effectiveness of our model. Additionally, the images demonstrate the time-varying nature of our attention mechanism, meaning the model switches its focus as new information arrives from the previous frames.

6. Conclusion

This paper introduces a CNN-RNN based model with a time-varying attention mechanism for interaction classification in multi-person sports videos. Our model can localize the key actors in the scene without requiring key actor annotations. Our method is flexible to the number of people in each frame and video, therefore applicable to any multi-person scenario. In this paper, we evaluate our model on the dataset of ice hockey penalties. Our dataset includes three interaction classes, ground truth pose, and hockey stick annotations for the players. We demonstrate the proposed model performance on our dataset and compare it against a few popular action recognition methods. We emphasize the effectiveness of hockey stick annotation through ablation studies and finally visualize the qualitative performance of the attention mechanism on some examples from our dataset.



(a) GT: slashing, Pred: slashing (b) GT: no penalty, Pred: no penalty (c) GT: slashing, Pred: no penalty (d) GT: tripping, Pred: no penalty

Figure 7. Visualization of attention mechanism on some examples. The figure demonstrates the link between the attention performance and the interaction classification.

References

[1] Cvat annotation tool. 4

[2] Icydata. 3

[3] National hockey league website. 3

[4] Zixi Cai, Helmut Neher, Kanav Vats, David A Clausi, and John Zelek. Temporal hockey action recognition via pose and optical flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 3

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[7] Lei Chen, Mengyao Zhai, and Greg Mori. Attending to distinctive moments: Weakly-supervised attention models for action localization in video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 328–336, 2017. 3

[8] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015. 3

[9] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7024–7033, 2018. 3

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 3, 6

[11] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 3

[12] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition, 2021. 6

[13] Mehrnaz Fani, Helmut Neher, David A Clausi, Alexander Wong, and John Zelek. Hockey action recognition via in-

tegrated stacked hourglass network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 29–37, 2017. 3

[14] Mehrnaz Fani, Pascale Berunelle Walters, David A Clausi, John Zelek, and Alexander Wong. Localization of ice-rink for broadcast hockey videos. *arXiv preprint arXiv:2104.10847*, 2021. 3

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 6

[16] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. R-cnns for pose estimation and action detection. *arXiv preprint arXiv:1406.5212*, 2014. 3

[17] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[19] Fa-Ting Hong, Wei-Hong Li, and Wei-Shi Zheng. Learning to detect important people in unlabelled images for semi-supervised important people detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4146–4154, 2020. 3

[20] Mihir Jain, Jan Van Gemert, Hervé Jégou, Patrick Bouthemy, and Cees GM Snoek. Action localization with tubelets from motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 740–747, 2014. 3

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 3

[22] Yanli Ji, Hong Cheng, Yali Zheng, and Haoxin Li. Learning contrastive feature distribution model for interaction recognition. *Journal of Visual Communication and Image Representation*, 33:340–349, 2015. 3

[23] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014. 3

[24] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[25] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Leveraging structural context models and ranking score fusion for human interaction prediction. *IEEE Transactions on Multimedia*, 20(7):1712–1723, 2017. 3

[26] Tian Lan, Yang Wang, and Greg Mori. Discriminative figure-centric models for joint action localization and recognition. In *2011 International conference on computer vision*, pages 2003–2010. IEEE, 2011. 2

[27] Tian Lan, Yang Wang, and Greg Mori. Discriminative figure-centric models for joint action localization and recognition. In *2011 International conference on computer vision*, pages 2003–2010. IEEE, 2011. 3

[28] Alessandro Prest, Vittorio Ferrari, and Cordelia Schmid. Explicit modeling of human-object interactions in realistic videos. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):835–848, 2012. 3

[29] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3043–3053, 2016. 2, 3, 4

[30] Michalis Raptis, Iasonas Kokkinos, and Stefano Soatto. Discovering discriminative action parts from mid-level video representations. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1242–1249. IEEE, 2012. 3

[31] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2657, 2013. 3

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jun 2017. 6

[33] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010. 1

[34] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 280–289, 2017. 3

[35] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1

[36] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 3

[37] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 3

[38] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015. 3

[39] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 3

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[41] Yicong Tian, Rahul Sukthankar, and Mubarak Shah. Spatiotemporal deformable part models for action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2642–2649, 2013. 3

[42] Moumita Roy Tora, Jianhui Chen, and James J Little. Classification of puck possession events in ice hockey. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 147–154. IEEE, 2017. 3

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[44] Arash Vahdat, Bo Gao, Mani Ranjbar, and Greg Mori. A discriminative key pose sequence model for recognizing human interactions. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1729–1736. IEEE, 2011. 3

[45] Kanav Vats, Mehrnaz Fani, David A Clausi, and John Zelek. Multi-task learning for jersey number recognition in ice hockey. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, pages 11–15, 2021. 3

[46] Kanav Vats, Mehrnaz Fani, David A Clausi, and John Zelek. Puck localization and multi-task event recognition in broadcast hockey videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4567–4575, 2021. 3

[47] Kanav Vats, William McNally, Chris Dulhanty, Zhong Qiu Lin, David A Clausi, and John Zelek. Pucknet: Estimating hockey puck location from broadcast video. *arXiv preprint arXiv:1912.05107*, 2019. 3

[48] Kanav Vats, William McNally, Pascale Walters, David A Clausi, and John S Zelek. Ice hockey player identification via transformers. *arXiv preprint arXiv:2111.11535*, 2021. 3

[49] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John Zelek. Player tracking and identification in ice hockey. *arXiv preprint arXiv:2110.03090*, 2021. 3

[50] Limin Wang, Yu Qiao, and Xiaoou Tang. Video action detection with relational dynamic-poselets. In *European conference on computer vision*, pages 565–580. Springer, 2014. 3

[51] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3

[52] Huimin Wu, Jie Shao, Xing Xu, Yanli Ji, Fumin Shen, and Heng Tao Shen. Recognition and detection of two-person interactive actions using automatically selected skeleton features. *IEEE Transactions on Human-Machine Systems*, 48(3):304–310, 2017. 3

[53] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 6

[54] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 3

[55] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012. 1

[56] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012. 3

[57] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 3

[58] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017. 3