Action Spotting

Harshit Monish
University at Buffalo
hmonish@buffalo.edu

Sidesh Shinde
University at Buffalo
sshinde5@buffalo.edu

Abstract

In team sports like football both person level information and group level contexts are crucial for event recognition. The task of spotting consists of finding the anchor time that identifies an event in a video and then classify the event to appropriate class. A common pipeline comprise of proposing temporal segments which are in turn further pruned and classified. Common methods for activity classification and detection make use of dense trajectories, actions' estimation, recurrent neural network, tubelets and handcrafted features. In order to recognize or detect activities within a video, a common practice consists of aggregating local features and pooling them, looking for a consensus of characteristics. In this project we focus our analysis on action spotting in soccer broadcast videos.

1 Introduction

Sports industry is a lucrative sector, with large amounts of money being invested on players and team. Even though the main scope of soccer broadcast is entertainment, such videos are also used by professionals to generate statistics, analyze strategies and scout new players. Automated sports video understanding can help in the localization of the salient actions of a game. Many recent methods exist to solve generic human activity localization in video focusing on sports, However detecting soccer actions is a difficult task due to the sparsity of the events within a video.

Action spotting involves pin-pointing any event that took place in the entire video/game. This helps in subsequent analysis in case of replay generation, statistical analysis of the game as well as the teams and players, etc. The task of action spotting can be challenging specially when there are multiple events that can resemble the same event. For example, the events of penalty and goal can both resemble a player shooting the ball at the goal post and other players around him. The imbalance of events occurring in the whole video further increases the difficulty of the task to spot events that have few occurrences. e.g events like clearance, ball out of play, throw-in have more number of occurrences whereas events like yellow card, red card have few occurrences. Hence, there is a semantic meaning for every action that needs to be involved in the classification task. In this project we leverage the annotations from SoccerNet-v2 [3] dataset which comprise of 17 classes of actions in a soccer game.

2 Related Work

Many automated sports analysis methods have been developed in the computer vision community to understand sports broadcasts. Early work used camera shot segmentation and classification to summarize games or focused on identifying video patterns to detect salient actions of the game [8]. Later Bayesian networks was used to detect goals, penalties, corner kicks and card events or to summarize games [5]. Recently deep learning approaches have been applied like Convolution Neural Network (CNN) for global descriptors extraction , Long Short-Term Memory networks [9] to temporally traverse soccer videos to identify the salient actions by aggregating temporally aggregating particular features [10]. The features can be global and local descriptors. Besides features, semantic information, such as player localization, as well as pixel information are also used to train attention models to extract relevant frame features. Some of the works propose to identify kicks, goals in soccer games using automatic multi-camera-based systems [6]. Another work uses logical rules to define complex events in soccer videos in order to perform visual reasoning on these events.

Differentiable pooling techniques have been applied in VLAD [6] which learns clusters of features descriptors and defines an aggregation of features. NetVLAD [1] generalizes VLAD by soften-

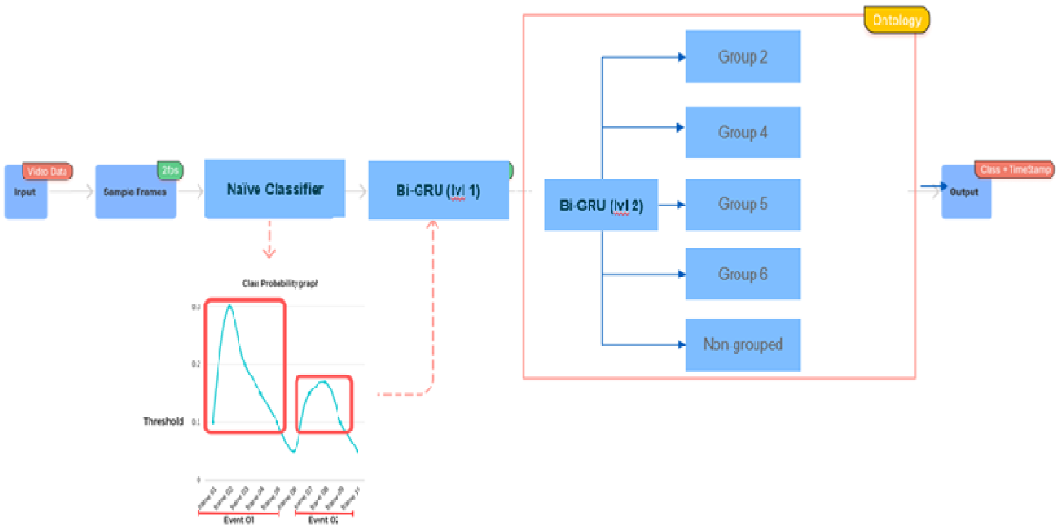

Figure 1: System Architecture

ing the assignment and disentangling the definition of the cluster and the assignment of the samples. Further NetVLAD++ [4] have implemented temporal aware pooling which consider the order of the frames also taking into account the past action event frames and post action event frames. It learns 2 different NetVLAD pooling modules for the frame features for before action event frames [-Tb, 0] and after action frames [0, Ta] and then aggregates the clusters information. CALF [2] address the task by developing a context aware loss for a temporal segmentation of module, and a YOLO-like [7] loss for an action spotting module. They first re-encoded the annotations and then compute the losses of frame segments based on the re-encodings. Transformer based temporal detection techniques [12] have also been implemented by baidu. In their implementation they have used the NetVLAD++ model and then have implemented a transformer encoder on top of it comprising of 3 encoding layers and the sine, cosine positional embeddings as in [11]

3 Implementation

The action spotting problem is 2-fold, i.e. it consists the time of the action and the category/class of the action. The proposed framework first classifies the timestamps of the actions in the entire game and then it classifies each of the timestamp. The system architecture is shown in Figure 1. The soccer game is given as input to the system. Feature generation is performed on the game-play. For the current approach, Baidu's feature-maps [12] are used to train the subsequent models. Once the feature-maps are obtained, the first regression task of predicting the time of events is done by the Naive Classifier as shown in Figure 2.

The Naive Classifier consists 2-layer deep neural network. The output of the classifier is binary (action, non-action) for each frame. When we pass all the frames from a game as an input to the Naive Classifier, the output would resemble the graph shown in Figure 2 with the frames on x-axis and their probability of being an action as y-axis. The final predicted time-stamps of the actions are the peaks from this 1-D signal generated bu the classifier.

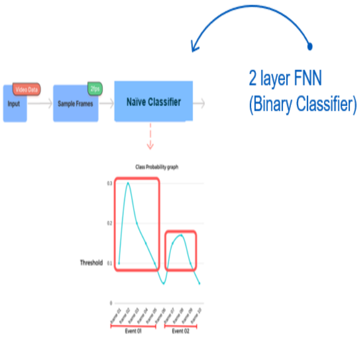

Figure 2: Naive Classifier

The predictions from the Naive Classifier are used to generate action windows. Considering the current frame as 0, t frames before and after the predicted frames are grouped together which forms the action window. t is set to 5 for this implementation. These windows are passed as an input to the level 1 Bi-GRU model. The level-1 Bi-GRU model consists 2 bidirectional GRU layers stacked on top of each other. The output of the model is a 17 node layer, giving the probability of an event using softmax activation. Based on the confidence of the level-1 model, the second level classification is decided. If the level-1 classifier predicts the action with less probability, the window is then passed to the level-2 classifier to again perform classification on it.

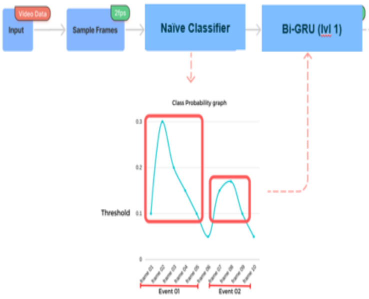

Figure 3: Level-1 Classifier

The level-2 classifier is same as the level-1 classifier with stacked bidirectional GRU layers. The key differences between the level-1 and level-2 classifier is that the level-2 classifier is trained specifically on pre/post frames of the predicted time of the action and also, it is trained on sub-group classification. Table 1 gives the classes in each group. These groups are formed on the potential actions that might seem similar to each other based on the exact time of the action. Since the frames before and after the action may capture different data, hence, they will help us to correctly classify these similar actions. Hence, if the level-1 classifier classifies an action from group x with a less probability/confidence, the pre/post frames of that timestamp are sent to the group x level-2 classifier to again predict the class. The level-2 classifier is trained only of the actions from their respective group, hence, allowing them to increase the classification accuracy on their respective classes.

For the level-2 classifier, t frames are used before/after the predicted frame to generate the action window for classification. Pre-action frames are used for groups 2, 5, and 6, whereas, post-action frames are used for group 4.

The number of frames for the level-1 and level-2 classifier along with the choice of either pre or post action frames are decided based on the ablation study.

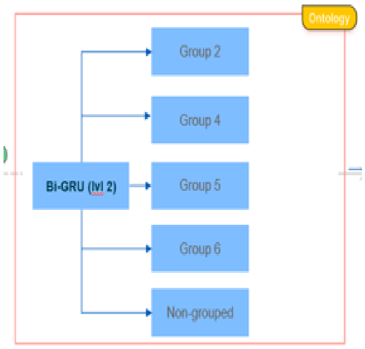

Figure 4: Level-2 Classifier

Group Number	Classes
Group 2	Goal, Shots on target, Shots off target, Corner
Group 4	Yellow card, Red card, Yellow->red card, Offside
Group 5	Clearance, Ball out of play, Throw-in
Group 6	Foul, Indirect free-kick, Direct free-kick

Table 1: Grouped classes for level-2 classifier.

4 Experimental Results

Dataset: The SoccerNet-v2 dataset comprise broadcast videos of total 550 games of total 764 hours with 720p and 224p resolutions at 25fps. We split the data into train set comprising of 300 games, test set of 100 games and validation set of 100 games. The challenge set comprise of 50 games. There are total of 110,458 annotated actions in the dataset on average of 221 actions per game, 1 action every 25 seconds. Each label is further marked as visible and non-visible. We have used the features provided by Baidu [13] which are extracted at 1fps.

Evaluation: We report the performance of our method using the Average-mAP metric introduced by SoccerNet-v2.A predicted action spot is positive if it falls within the given tolerance x of a ground-truth timestamp from the same class. The Average Precision (AP) based on PR curves is computed and then averaged over the classes(mAP), after which the Average-mAP is the AUC of the mAP at differ-

Model	Test	Challenge
Baidu	47.05	49.56
NetVLAD++ with Baidu features	NA	43.99
Our Implementation	**35.17**	**36.71**
AImageLab	28.83	27.69
CALF$_{Calibration}$	NA	15.83
CALF	NA	15.33
NEtVLAD++	11.51	9.91
NEtVLAD	4.20	4.31

Table 2: Average-mAP on tight bound (5sec)

Model	Test	Challenge
Baidu	73.77	78.84
NetVLAD++ with Baidu features	NA	74.63
NEtVLAD++	53.4	52.54
Our Implementation	**49.86**	**51.35**
CALF$_{Calibration}$	46.80	46.39
AImageLab	28.83	27.69
CALF	NA	42.22
NEtVLAD	31.37	30.74

Table 3: Average-mAP on loose bound (60 sec)

ent tolerances x. The tolerance values are 5sec for tight bounds and 60 sec for loose bound.

Results: For tight bound of 5sec our model gave mAP of 35.17 on test data and 36.71 on challenge set as shown in Table 2. On loose bounds our model gave mAP 0f 49.86 on test and 51.35 on challenge set as shown in Table 3. Further we have shown the average mAP per class results in Table 4. The average-mAP for classes that have less occurrences have less mAP value because of imbalance in the dataset whereas for the classes that have higher occurrences have high mAP value.

5 Conclusion

In this project we have implemented a multimodal architecture for action spotting task in soccer videos using Soccernet-v2 dataset.The experimental results shows that the proposed scheme can incorporate pre-event and post-event frames for classification task, especially for closely related events. The reason for Red Card and Yellow->red card event's poor accuracy is because of the imbalance in the dataset. The reason for Shot on Target and Shot off Target event's poor accuracy is because of the closeness with Goal event. Further improvement in performance can be obtained by exploring different solutions for sparsity problem and incorporating vision transformer for action spotting.

Class	mAP
Penalty	88.0
Kick-off	55.0
Goal	58.0
Substitution	59.0
Offside	52.0
Shots on target	14.0
Shots off target	33.0
Clearance	53.0
Ball out of play	53.0
Throw-in	67.0
Foul	65.0
Indirect free-kick	45.0
Direct free-kick	45.0
Corner	69.0
Yellow card	57.0
Red card	9.0
Yellow->Red card	5.0

Table 4: Average-mAP on Test Dataset for Each class

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2020.

[3] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4508–4519, 2021.

[4] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2021.

[5] Chung-Lin Huang, Huang-Chia Shih, and Chung-Yuan Chao. Semantic analysis of soccer video using dynamic bayesian network. *IEEE Transactions on Multimedia*, 8(4):749–760, 2006.

[6] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[8] Reede Ren and Joemon M Jose. Football video segmentation based on video production strategy. In *European Conference on Information Retrieval*, pages 433–446. Springer, 2005.

[9] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[10] Takamasa Tsunoda, Yasuhiro Komori, Masakazu Matsugu, and Tatsuya Harada. Football action recognition using hierarchical lstm. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 99–107, 2017.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *arXiv preprint arXiv:2106.14447*, 2021.

[13] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection, 2021.