EVENTS RECOGNITION FOR A SEMI-AUTOMATIC ANNOTATION OF SOCCER VIDEOS: A STUDY BASED DEEP LEARNING

L.F. Kazi Tani [1],*, A. Ghomari [1], M.Y. Kazi Tani [2]

[1] RIIR Laboratory, Computer Science Department, Exact Sciences and Applied Faculty, University of Oran1 Ahmed Ben Bella, Oran, Algeria - (lamiakazitani, ghomari65)@yahoo.fr
[2] LabRI-SBA Lab., Ecole Superieure en Informatique, Sidi Bel Abbes, Algeria - yassine.kazi@gmail.com

KEY WORDS: Soccer Video Annotation, Event detection, Deep Learning, CNN, RNN, Mask R-CNN.

ABSTRACT:

In this work, we propose an efficient way of web video annotation in soccer domain. To achieve this, it is necessary to enjoy different architectures of deep learning. We aim at realising a system of annotation able to recognise several events from information of the object that is the ball in our case, in order to fuse them as a part of actions in video. We propose to use Deep Neural Network (DNN) to detect ball and actions. However, Mask R-CNN can play a very important role for features extracted as an output using a training network on ImageNet dataset. The Mask R-CNN is chosen as a method using different CNN as backbone (convolutional Neural Network) ResNet50, ResNet101 and ResNet152, VGG16, VGG 19. We experimentally verify the effectiveness of the proposed method in the test phase.

1. INTRODUCTION

The huge collection of videos on the web, form a set of tasks in computer vision research for annotation, indexing and retrieval such as summarisation, actions recognition, events detection, object tracking... In Artificial intelligence, recognising human actions is as important as constructing models around, for interpretation. From Web images and videos, we have a lot of an unstructured data, shared, saved and published through the internet. Therefore, we assume that annotation is crucial for indexing web videos. However, in sport domain, it is usually controlled by a set of rules or customs, which serve to ensure fair competition, and allow consistent adjudication of the winner. Hence, from images of these videos, we have to retrieve not only the image itself or the video, but exactly all interesting objects and actions forming events. This can be called as recognition where events can be considered in general as real world happenings yet attended by people, and where the related multimedia is captured by the attendees (A. Kashif et al. 2018).
For our study, we consider the domain of soccer. So, it is a team sport played between two teams of eleven players with a spherical ball. It is played by million players all around the world, making it the world's most popular sport (Association football, 2018). However, rules must be applied at the moment of a new detected action. FIFA sets up 17 laws in the official Laws of the Game, each containing a collection of stipulation and guidelines. So, events look like yellow card, red card, foul, substitution, goal, see also, own goal, corner kick, hand ball, offside, penalty kick,…. Here, researchers have to distinguish studies about object recognition and event recognition. We proceed on two steps, first, on the object that is the ball and action recognition, and second, on event recognition. Here recognition of these events is crucial. Hence, this is possible via an annotation system able to agree what type of event is or will be. Different actions in soccer exist like goal, corner, foul, yellow and red cards… where every user is interested by skimming the entire video to view a particular event. We describe an event as an action detected along duration with a beginning and an end time. In this paper, we propose novel annotation system based events recognition from soccer web videos.

Our study is based Deep Learning, we are based Convolutional Neural Network (CNN) for the recognition of the ball in images, and on the other hand, Recurrent Neural Network (RNN) for the recognition of the event from an action detection; both are the two most widely used deep neural network models; they are able of learning right spatial and temporal features, respectively (Bai et al., 2018). However, For RNN, we choose the Mask-RCNN method for our event recognition (He et al., 2017). .
Actually, Deep Learning proves its efficiency in all domains. In 1980, Fukushima (Fukushima, 1980) took patches of all layers and applied neurons with the same parameters on these patches for the next layers to obtain the translational invariance, thus this can be considered as the predecessor of the Convolutional Neural Network CNN. At the end of the 90s, CNN has revolution the research area in computer vision, where it consists of an input and an output layer, as well as multiple hidden layers. Whereas, by browsing videos from the web, and for events detection in soccer videos (Yu J., et al., 2019), we have to extract the spatio-temporal features and to classify them in several classes as different events. In this paper, we aim to annotate events from soccer videos downloaded from YOUTUBE.

An action in soccer is a reply to a change of the game from both player and ball causing the beginning of an event on time. As said in (G Yao et al., 2019), we have two ways for representation of action recognition; the Handcrafted representation method (Caba Heilbron et al., 2016, Mettes et al., 2015, Yu, 2015), and the Deep Learning representation method (I. Goodfellow, 2016); in the first method, we extract features manually and it is generally used as a baseline to evaluate new Deep Learning representation; whereas, the deep learning representation method learns the trainable features automatically from videos (G Yao et al., 2019). So, talking about automatic human events recognition recently seems that researchers in this area aim to go beyond the human spirit.

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W16-135-2019

In this way, work in (Feichtenhofer et al., 2016, Simonyan, et al., 2014, Ramanathan, 2015, Gan, , 2015, Jain, 2015, Tang, 2012, Wang,, 2016c), focus on modelling the motion information and temporal dynamics. Authors in (K Simonyan et al., 2014) developed a two-stream CNN for action recognition, where one, captured the static appearance and the other, described the temporal motion information. Also, (M. Jain et al., 2015) integrated object responses with motion features for action recognition in videos, where (V. Ramanathan et al, 2015) proposed a new temporal embedding method to capture video structure for event retrieval and recognition. Authors in (L. Wang, 2016c) modelled long-term temporal structure with a sparse sampling strategy and temporal aggregation module, which obtained good performance on the standard action recognition benchmarks. In parallel, work on events recognition in semantic representations of high level concepts was been studied in (Ebadollahi et al., 2006, Liu, et al., 2013), (Mazloom, M. et al., 2014), (Habibian, A. et al., 2014, Izadinia, et al., 2012). Hence, (Mazloom, M. et al., 2014) proposed a new algorithm to learn what concepts are most informative per event, called as conceptlets, they solved this problem with an importance sampling method (Wang et al., 2017). In (Chaquet et al., 2013), (Hassner et al., 2013) and (Giancola, et al., 2018), authors provided reviews of datasets and benchmarks designed for events recognition. Since 1998, the Convolutional Neural Network CNN makes the buzz in the domain of image and video (Y. Lecun et al., 1998). Where, authors in (Dawn et al., 2016) and (Zhen et al., 2016) present local representation methods for event recognition. However, researchers in (F. Zhu et al., 2016) discuss superiorities and limitations of existing methods. During these 5 last years, the CNN improve the performance of image classification, traking, object detection, event recognition. Thereafter, several architectures had been proposed for visual recognition tasks since LeNet-5, they Worked and trained the framework of CNN. AlexNet, GoogleNet (Jefferson Ryan et al., 2016), VGGNet (A. Diba et al., 2016), ResNet (Lecun et al., 2014), where this one is the latest deep architectures, has remarkably increased the depth of CNN to 152 layers using deep residual layers with skip connections. ResNet has recently been extended to more than 1000 layers on the CIFAR-10 dataset (Dawn et al., 2016). So, CNN imposes himself on the community of research in images and videos by improving performance of the results obtained from diverse methodologies and architectures of the CNN (Gan, et al., 2015).

The rest of this article is organise as follow: in section 2 we present relevant works to events detection from action recognition in soccer domain; in section 3, we give our method on how to procced from an input video of soccer to an output event recognition; in setion 4, we highlight the performance of the method by experiments, and in the end section, we conclude and discuss the solution.

2. BACKGROUND AND RELATED WORKS

This section focus on relevant works related to events detection from videos and especially in soccer. Basically, there are 17 laws in soccer, each containing a collection of stipulation and guidelines published by FIFA, but are maintained by the International Football Association Board (IFAB). Therefore, broadcast videos are freely available from the Internet (e.g. YouTube, Daily Motion, Vimeo…). Then, our interest consists to develop a system based on events detection in soccer from web videos. Here, (Kolekar et al., 2015) use audio features to detect events in soccer scenes and generate highlights. Also, (Sukhwani et al., 2017) produces semantically meaningful and compelling summaries by generating highly personalized variable length. (Zawbaa, 2012) Presents an approach for automatic soccer videos summarization using machine learning techniques. (J. Chen et al., 2019) proposes a framework for sports camera selection using web videos to address the data scarcity problem. Also, (Bhaumik et al., 2016), (Ramanathan et al., 2016) present a method able to illustrate that the quality of the generated summary is not degraded by removing duplicate frames having nearly the same visual content, where, an additional contribution here, is the determination of an automatic threshold for elimination of redundant frames. Hybrid summarisation techniques attempt to achieve enhanced levels of semantic summarisation by combining the analysis of internal and external information (Xu et al., 2006). Although, (Huang et al., 2006) interprets semantic information to detect events automatically ensured by (Automatic P. et al., 2015) who define the output summary as an aggregate of clips containing a set of interest events from a user perspective. Even in Tennis, authors in (Ghosh et al., 2017) propose a score recognition algorithm to index all the video segments with their scores to create a navigable and searchable match.

Furthermore, with the emergence of Deep Learning, several works aim to present objects or events detection in soccer videos. (Donahue et al., 2015) use a deep sequence model, which learn a deep hierarchy of parameters only in the visual domain, take a fixed visual representation of the input and only learn the dynamics of the output sequence, when (Baccouche et al., 2010) are based on RNN, they use such way to train a Long Short Term Memory (LSTM) (Hochreiter et al., 1997) network that temporally browses the video to detect events. In (M. Jefferson Ryan et al., 2016), (M. Shugao et al., 2016) an another concept involves the generation of dynamic image through the collapse of multiple video frame and the use of 2D deep feature exaction on such representations (Hakan et al., 2016). (Liu et al., 2017,Hong et al., 2018) Use CNN for event detection based temporal information. (Tsagkatakis et al., 2017) present a two-stream approach to detect goals, while (R. A. Sharma et al., 2018) present a method to compute projective transformation between a static model and a broadcast image as a nearest neighbour search and show that the approach gives highly accurate results.

Basically, CNN was designed to extract 2D spatial features from still image while videos are considered as 3D spatiotemporal signals whereas, the core issue to extend CNN from images to videos is the temporal information exploitation (Voulodimos et al., 2018). In order to model spatial (RGB frame) and temporal information (optical flow) separately and average the predictions in the last few layers of the network, (Simonyan et al., 2014), (Feichtenhofer et al., 2016) propose another set of video analysis architecture for spatiotemporal fusion called the Two-stream networks. Consequently, several researchers team propose to modify and classify temporal segments (Gao et al., 2018, Shou et al., 2017, Gao et al., 2017, Zhao et al., 2017, Sigurdsson et al., 2017, Jain et al., 2015) Present a method for zero shot action recognition without using any video examples. In fact, performances on standard benchmarks such as UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al, 2011) are achieved by a combination of ConvNets and a Fisher Vector encoding (Izadinia, et al., 2012) of hand-crafted features over dense trajectories (H. Wang et al., 2013). Also, in literature, work in (Shou et al., 2017) design a novel Convolutional De-Convolutional (CDC) network that places CDC filters on the top of 3D ConvNets, this is effective for abstracting action semantics but reduce the temporal length of the input data. Recently, (Diba et al., 2016) combines the 3D convolutions and the two-stream approach for video classification, achieving state-of-the-art performance at significantly lower processing times.

3. METHOD

Our interest aims to improve the annotation systems in the domain of soccer videos. The idea is to detect events from object that is the ball and the actions will be deduced after. Indeed, authors in (Zhu et al., 2016) divide the research field in soccer into three categories: video summarization, augmentation and high-level features (Figure1.)

The video summarization comprises detection of important moments in a match browsing all the video to create summaries; Augmentation includes the field of improving spectator's experience; and, The high-level features concern tracking of the ball and actions detection, analysis of team statistical and real time event.

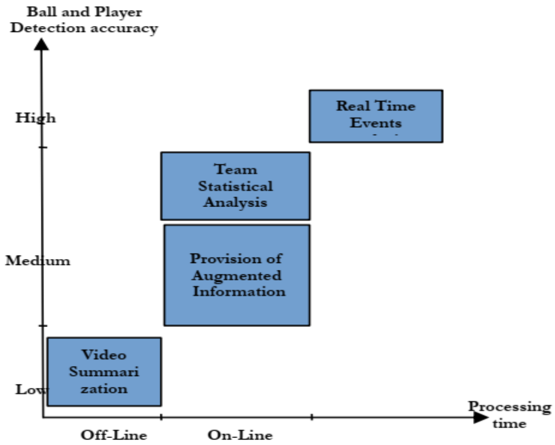

Figure1. The three categories of Soccer Fields.

Our area of research concerns the third category from (Zhu et al., 2016) to satisfy annotation of soccer videos from the web. We are inspired from two works (Liu et al., 2017,Yu et al., 2019) the first used deep convolutional neural network to extract features from video frames and calculated Euclidean distance to measure the difference between the consecutive frames; and the second is a multiple temporal scales and story generation using the relationship between events and replays in soccer videos These two methods are based on the replay for the detection of events.

3.1 Pre-Processing

We consider that the position of the ball is the basis to study the event detection in soccer videos. Indeed, tracking the ball over multiple frames can be used to see if the existence of the ball is detected. Now, adding detection of players enhance counting passes. Then, tracking of spatio-temporal data (3D) is possible since it concerns the ball.

However, CNN is a special type of deep neural network that is suited to analyse visual imagery. This is essentially how the CNN works:

1. convolution allows us to extract appropriate features from the input videos (frames);
2. Pooling then allows us to reduce dimensionality of the feature maps but keep the most important information;
3. Forming fully connected layers then allows us to ensure connections to all activations in the previous layer.

Afterwards, we are interested in our study to 4 events in soccer: goal, penalty, offside and throw-in. (Figure 2)

1. The goal consists of defending team when the ball is goes directly into the opposing goal, having last been touched by the attacking team, from a situation in which the laws do not permit an attacking goal to be scored directly.
2. The penalty consists of taking a single shot on the goal while it is defended only by the opposing team's goalkeeper.
3. The offside consists of two conditions:
   a) Any part of the player's head, body or feet is in the opponents' half of the field (excluding the half-way line); and
   b) Any part of the player's head, body or feet is closer to the opponents' goal line than both the ball and the second-last opponent.
4. The throw-in consists of restarting play in soccer where the ball crossed the touch-line; it is taken by the opponents of the player who last touched the ball when it crossed the touch-line.

Therefore, we apply Mask-RCNN as a method to achieve recognition of events. We see that the goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyse. Segmentation can enhance annotation process once object is detected. We begin first, by features extraction using different architectures as a convolutional backbone. Then a network head for bounding-box recognition where for each RoI (Region of Interest), we apply mask prediction (K. He et al., 2017).

3.2 Features Extraction

Saying that the aim of our study is the annotation of soccer events, and considering that event detection is highly related with object and action recognition, in this paper, we need two distinct sub-networks for annotation: one for object and the second is for action recognition. We fine-tune each sub-network for event detection on our soccer event datasets and fuse these two sub-networks streams by combining the prediction scores of two separate sub-networks. However, to study deep features for our soccer ball and action recognition, as a backbone, we rely on the most particulars architectures of a pre-trained network in this field, VGGNet (Simonyan et al., 2014), and ResNet (He et al., 2016). Furthermore, VGGNet and ResNet are available in different configurations. Particularly, our study focuses on VGGNet16 and VGGNet19 with 16 and 19 layers respectively, whereas ResNet respects evaluation of the most configurations containing 50, 101, and 152 layers.

To satisfy features extraction, we use PyTorch 1.0.[1] as a framework from the last fully connected layer for both CNNs: VGGNet16, VGGNet19 and ResNet (i.e., Fc7 for VGGNet; Fc1000 for all configurations of ResNet). Mainly, we extract 10 features vectors through different architectures for each image.

[1] https://pytorch.org/

VGGNet returned a feature vector of size 4096; where ResNet (with all configurations) provided feature vectors of size 1000.

Considering that an event has a start and an end and therefore has duration, and particularly, from success of ResNet architecture, first, we use the pre-trained ImageNet-ResNet50, ResNet101 and Resnet152. Then, we fine-tune the model parameters for event detection on the training soccer event dataset. In addition, for action network, we resort advances on the task of action recognition. Notably, the ResNet-152 architecture includes a sequence of 50 residual blocks each comprising three convolutional layers followed by a batch-normalization layer and ReLU activation. The output of the third convolutional layer is added to the input of the residual block to produce the layer's output. We trained the network using a fully connected output layer with sigmoid activations. This focused on single-task recognition, using shared weights. The loss function optimized was the sigmoid cross-entropy.

Unlike image detection problems, feature extraction in video must address the challenges associated with the variation of the duration of events. Since soccer event requires a begin and an end time, it is fundamental that the current time interval depends on neighbouring time intervals, hence, to capture this temporal information we focus also on an RNN-based approach, because it performs better at sequence modelling tasks which requires flexible modelling of context dependencies (Cui J. et al., 2018). In addition, we gathered two different types of information for ball recognition from the CNN: ball binary presence from the output layer and ball features from the last pooling layer. The aim of training on ball features was to capture information (position of the ball) and visual cues (colour for example) that could potentially enhance event recognition.

4. EXPERIMENTS

4.1 Our Dataset (Table. 1)

This section introduces details of our experiments and analyse the results.

Results Events	Number of videos containing events	Images with events	Videos without events	images without events
Goal	350	2000	73	373
Penalty	244	1500	61	261
Offside	225	800	24	156
Throw-in	181	700	42	210

Table 1. Our dataset (downloaded videos and images)

In our experiments, we collect a training dataset from YOUTUBE to detect soccer video events of World cup Championship 2018. Therefore, from the web and from Google's Open Images Dataset V4[2] (Figure 2), we download images. These images have bounding-box annotations describing ball and actions. However, we collect images of only close-up camera shot, 3000 soccer ball images and 5000 soccer images for goal, penalty, offside, and throw-in events to be used for training. For each soccer event, the training data consists of an average of 2000 images.

Figure 2. Examples of ball, goal, penalty, offside and throw-in.

Also, we download from YOUTUBE 30 matches. Here, we produced a bunch of bulk screenshots using VLC player, and we kept images that we are only interested in.

Furthermore, we collect over 500 (cf.200) sequences of videos of 2-3 seconds containing events (called "nothing" without events, respectively). Videos called nothing represent no event, showing diverse examples of false events. Given an unknown soccer video as input, each video frame can be detected as belonging to any of these events categories or as nothing I.e no events. For images without events, we choose only 20% of the downloaded images containing events.

Note that we choose 60% from the download for training and 40% for the test.

4.2 Implementation

In each training batch, we randomly select a set of input images with a size of 32 × 32. 500 iterations of back-propagation create an epoch. To implement the proposed model we used Pytorch, the deep learning open-source library. Python 3.6 is used for all the implementations. All the implementations of the network are conducted on a workstation equipped with an Intel i7-6850K CPU with a 64 GB Ram and a single NVIDIA GTX Geforce 1080 Ti GPU and the operating system is Ubuntu 18.04.2. The learning rate is initialized to 0.02 for all layers and divided by 20 every 10 epochs.

Training our networks had taken more than 34 days for all backbones.

4.3 Results

As a result, we show in table 2, outputs of our study. Note that the *Precision* means the relation between true positives and all elements considered as positives. The *Recall* measures how good all the positives are found.

[2] https://storage.googleapis.com/openimages/web/index.html

Method	Backbone	Precision	Recall	mAP
Mask R-CNN	ResNet 50	0.799	0.681	0.542
	ResNet 101	**0.875**	0.713	**0.650**
	ResNet 152	0.860	**0.720**	0.510
	VGG 16	0.781	0.665	0.485
	VGG 19	0.790	0.646	0.473

Table 2. Interpretation overview.

We can say from the overview in Table 2 of the precision, the recall and the mAP that Mask R-CNN with backbone architecture ResNet 101 shows the best result in precision and mAP, in parallel for the recall the architecture with ResNet 152 gives the best results compared to the other.

From this study, we confirm that Mask R-CNN using a backbone like ResNet as an architecture can be considered as a success like an effective framework.

Our aim is to reuse a model that gives the best result in precision, so, form this study, we show that using Resnet-101 will be applied for our future study.

To compare our study to others in state-of-the-art, we'll be based on benchmarks on the future to approve our proposal.

5. CONCLUSION

Our study is based on Mask R-CNN method to design a system that is effective for generating annotations of 4 different events in soccer. We collected a dataset downloaded from YOUTUBE and consist of matches from FIFA World Cup 2018. We apply different architectures of Deep Learning as a backbone of this method (Mask R-CNN) where we achieved in-depth interpretation on our dataset and observed how these different architectures affect the performance of the network. We also see from the measurement that ResNet gives the best result as a backbone of Mask R-CNN. Future works will include collecting more videos and images finding more events of interest to improve the effectiveness of the network .We propose to follow the study (Kazi Tani et al., 2019) as a complete solution to this one. We perform ontology to infer the proper event. Note that we can use the extracted features as an input to our ontology. Hence, we can apply SWRL rules to prevent different events in soccer.

REFERENCES

Diba, A., Pazandeh, A.M., Van Gool, L., 2016. Efficient Two-Stream Motion and Appearance 3D CNNs for Video Classification. *arXiv preprint arXiv*:1608.08851, 2016.

Sigurdsson, G., Russakovsky, O., Gupta, A., 2017. What Actions are needed for Understanding Human Actions in Videos? *The IEEE International Conference on Computer Vision (ICCV)*, 2137-2146.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.*, 2018.

Association football. 2018. "https://en.wikipedia.org/wiki/Association_football#ci te_note-EB-6", *consulted on December 5th*, 2018.

Piamsa-nga, P., Bojukrapan, S., 2015. Automatic soccer archive summarization using time constraint. *International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai, 2015, 1-6. doi:10.1109/ICSEC.2015.7401455. 2015.

Hakan, B., Fernando, B., Gavves, E., Vedaldi, A., Gould, S., 2016. Dynamic image networks for action recognition. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 3034-3042.

Bhaumik, H., Bhattacharyya, S., Chakraborty, S., 2016. Redundancy Elimination in Video Summarization. *Studies in Computational Intelligence,* 630. Springer, Cham Elimination in Video Summarization. 2016.

Feichtenhofer, C., Pinz, A., Zisserman, A., 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1933-1941.

Xu, C., Wang, J., Wan, K., Li, Y., Duan, L., 2006. Live Sports Detection Based on Broadcast Video and Webcasting Text, in *Proceedings of the 14th annual ACM international conference on Multimedia* (MM '06). Santa Barbara, CA: ACM Press, 2006.

Huang, C.-L., Shih, H.-C., Chao, C.-Y., 2006. Semantic analysis of soccer video using dynamic bayesian network. *IEEE Transactions on Multimedia*, 8(4), 749–760.

Cui, J., Long, J., Min, E., Liu, Q., Li, Q., 2018. Comparative Study of CNN and RNN for Deep Learning Based Intrusion Detection System. *Lecture Notes in Computer Science,* 11067. Springer, Cham. 2018.

Dawn, D., Shaikh, S., 2016. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector, *Visual Computer*., 32 (3), 289-306.

Ebadollahi, S., Xie, L., Chang, S., Smith, J.R., 2006. Visual event detection using multi-dimensional concept dynamics. *In ICME*, 881– 884.

Caba Heilbron, F., Carlos Niebles, J., Ghanem, B., 2016. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1914–1923.

Zhu, F., Shao, L., Xie, J., Fang, Y., 2016. From handcrafted to learned representations for human action recognition, *Image and Vision Computing*., 55, 42-52.

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*.1980.
Yao, G, Lei, T., Zhong, J., 2019. A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognition Letters*, 118, 14-22.

Tsagkatakis, G., Jaber, M., Tsakalides, P., 2017. Goal!! event detection in sports video. *Electronic Imaging*, 2017(16), 15–20.

Yu, G., Yuan, J., 2015. Fast action proposals for human action detection and search. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 1302–1311.

Gan, C., Wang, N., Yang, Y., Yeung, D., Hauptmann, A.G., 2015. Devnet: A deep event network for multimedia event detection and evidence recounting. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2568–2577.

Ghosh, A., Jawahar, C.V., 2018 SmartTennisTV: Automatic Indexing of Tennis Videos. In: *Rameshan R., Arora C., Dutta Roy S. (eds) Computer Vision, Pattern Recognition, Image Processing, and Graphics.* 2018.

Giancola, S., Amine, M., Dghaily, T., Ghanem, B., 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In: *CVPR* Workshops. 2018.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. HMDB: a large video database for human motion recognition. In: *ICCV*, 2556-2563.

Zawbaa, H.M., El-Bendary, N., Hassanien, A.E., Kim, T.-h., 2012. Event Detection Based Approach for Soccer Video Summarization Using Machine learning, *International Journal of Multimedia and Ubiquitous Engineering*, 7(2), 63-80.

Wang, H., Schmid, C., 2013. Action recognition with improved trajectories. In Proc. *ICCV.*

Habibian, A., Snoek, C.G.M., 2014. Recommendations for recognizing video events by concept vocabularies. *Computer Vision and Image Understanding*, 124, 110–122.

Goodfellow, I., Bengio, Y., A. Courville. Deep Learning. MIT Press, *MIT Press*. 2016.

Izadinia, H., Shah, M., 2012. Recognizing complex events using large margin joint low-level event model. *In 12th European Conference on Computer Vision*, IV, 430–444.

Chaquet, J., Carmona, E., Fernández-Caballero, A., 2013. A survey of video datasets for human action and activity recognition, *Computer Vision and Image Understanding*, 117 (6), 633-659.

Chen, J., Lu, K., Tian, S., Little, J., 2019. Learning Sports Camera Selection From Internet Videos, *2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village*, HI, USA, 2019, 1682-1691. doi:10.1109/WACV.2019.00184. 2019.

Donahue, J. Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long -term Recurrent Convolutional Networks for Visual Recognition and Description. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2625-2634.

Gao, J. Chen, K., Nevatia. R., 2018. CTAP: Complementary Temporal Action Proposal Generation. In: *The European Conference on Computer Vision (ECCV)*, 68-83.

Gao, J., Yang, Z., Chen, K., Sun, C., Nevatia, R., 2017. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In: *IEEE International Conference on Computer Vision (ICCV)*, 3628-3636.

Jain, M., Van Gemert, J.C., Snoek, C.G., 2015. What do 15,000 object categories tell us about classifying and localizing actions?. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 46-55.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. *´ arXiv:1703.06870*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *CoRR, abs*/1409.1556.

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

Ahmad, K., Mekhalfi, M.L., Conci, N., Melgani, F., Natale, F.D., 2018. Ensemble of deep models for event recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2), 51.

Kingma, D.P., Ba, J., Adam: A method for chastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Wang, L., Wang, Z., Qiao, Y., Van Gool, L., 2017. Transferring deep object and scene representations for event recognition in still images. *IJCV*, 2017.

Kazi Tani, L.F., Ghomari, A., Kazi Tani, M.Y., 2019. A Semi-Automatic Soccer Video Annotation System based on Ontology Paradigm. *Paper accepted at ICICS19*, Irbid. Jordan. 2019.

Liu T. et al. Soccer Video Event Detection Using 3D Convolutional Networks and Shot Boundary Detection via Deep Feature Distance. In: *Liu D., Xie S., Li Y., Zhao D., El-Alfy ES. (eds) Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science*, vol 10635. Springer, ChamVideo. 2017.

Liu, J., Yu, Q., Javed, O., Ali, S., Tamrakar, A., Divakaran, A., Cheng, H., Sawhney, H.S., 2013. Video event recognition using concept attributes. In: *WACV*, 339–346.

Sukhwani, M., Kothari, R., 2017. A Parameterized Approach to Personalized Variable Length Summarization of Soccer Matches. *arXiv preprint arXiv:1706.09193*.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A., 2010. Action classification in soccer videos with long short-term memory recurrent neural networks. *Artificial Neural Networks–ICANN 2010*, 154–159.

Kolekar, M.H., Sengupta, S., 2015. Bayesian Network-Based Customized Highlight Generation for Broadcast Soccer Videos, *IEEE Transactions on Broadcasting*, 61(2), 195-209.

Jain, M., van Gemert, J.C., Mensink, T.S., Cees, G., Snoek, M., 2015, Objects2action: Classifying and localizing actions without any video example. *The IEEE International Conference on Computer Vision (ICCV),* 4588-4596.

M. Jefferson Ryan, and A. Savakis. Anomaly Detection in Video Using Predictive Convolutional Long Short-Term Memory Networks. *arXiv preprint arXiv:1612.00390*, 2016.

M. Shugao, L. Sigal, and S. Sclaroff. ”Learning activity progression in lstms for activity detection and early detection.” In Proceedings of the IEEE *Conference on Computer Vision and Pattern Recognition*, pp. 1942-1950, 2016.

Mazloom, M., Gavves, E., & Snoek, C. G. M. Conceptlets: Selective semantics for classifying video events. *IEEE Transactions on Multimedia*, 16(8), 2214–2228. 2014.

Mettes, P., van Gemert, J. C., Cappallo, S., Mensink, T., Snoek, C.G., 2015. Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting. *In Proceedings of the 5th ACM on ICMR*, pages 427–434.

Sharma, R.A., Bhat, B., Gandhi, V., Jawahar, C.V., 2018, Automated Top View Registration of Broadcast Football Videos. 208 *IEEE Winter Conference on Applications of Computer Vision (WACV),* Lake Tahoe, NV, 2018, 305-313.

Ramanathan, V., Tang, K.D., Mori, G., Li, F., 2015, Learning temporal embeddings for complex video analysis. In *ICCV*, 4471–4479.

Bai, S., Kolter, J. Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271.*

Hochreiter, S., Schmidhuber. J., Long short-term memory. Neural Computation, 1997.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. *In NIPS*, 568–576.
T. Hassner, 2013. A critical review of action recognition benchmarks, In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 245-250.

Tang, K. D., Li, F., Koller, D., 2012. Learning latent temporal structure for complex event detection. In *CVPR*, 1250–1257.

Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., Fei-Fei, L., 2016. Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3043–3053.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 20–36.

Zhen, X., Shao, L., 2016. Action recognition via spatio-temporal local features: A comprehensive study. *Image and Vision Computing*. 50 1-13.

Hong, Y., Ling, C., Ye. Z., 2018. End-to-end soccer video scene and event classification with deep transfer learning. 2018 *International Conference on Intelligent Systems and Computer Vision* (ISCV), Fez, 2018, 1-4.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition, in: *Proceedings of the IEEE,* 2278- 2324.

Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D., 2017. Temporal Action Detection with Structured Segment Networks. *The IEEE International Conference on Computer Vision (ICCV)*, 2914-2923.

Yu, J., Lei, A., Hu, Y., 2019. Soccer Video Event Detection Based on Deep Learning. In: *Kompatsiaris I., Huet B., Mezaris V., Gurrin C., Cheng WH., Vrochidis S. (eds) MultiMedia Modeling. MMM 2019. Lecture Notes in Computer Science, vol 11296*. Springer, Cham..

Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang. S., 2017. CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5734-5743.

https://doi.org/10.5194/isprs-archives-XLII-2-W16-135-2019