SOCCER EVENT DETECTION

Abdullah Khan[1,2], Beatrice Lazzerini[2], Gaetano Calabrese[3] and Luciano Serafini[3]

[1]Department of Information Engineering, University of Pisa, Pisa, Italy
[2]Department of Information Engineering, University of Florence, Florence, Italy
[3]Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

The research community is interested in developing automatic systems for the detection of events in video. This is particularly important in the field of sports data analytics. This paper presents an approach for identifying major complex events in soccer videos, starting from object detection and spatial relations between objects. The proposed framework, firstly, detects objects from each single video frame providing a set of candidate objects with associated confidence scores. The event detection system, then, detects events by means of rules which are based on temporal and logical combinations of the detected objects and their relative distances. The effectiveness of the framework is preliminary demonstrated over different events like "Ball possession" and "Kicking the ball".

KEYWORDS

Event detection in video, simple events, complex events.

1. INTRODUCTION

Identifying intermediate and high-level complex events from an unstructured video is an extremely challenging task due to the variation and the dynamics of the video sequence. In this work, the focus is on the analysis of videos showing team sport activities and, more specifically, soccer game. Given the nature of the game itself, where two teams each of eleven players produce a vast number of possible interactions, soccer is a highly complex system [16]. Due to the high complexity governing the "beautiful game", the statistical analysis of soccer games has fascinated scientists and experts.

Data are playing an increasingly key role in sports, but they must be processed to extract meaningful information [2, 3]. Data-driven decision plays a significant role in soccer and many other sports. Collecting and properly handling quality data from a soccer match is, therefore, clearly of immense value for a team, management and other stakeholders.

The data typically collected from a soccer game include: goals scored, assists, number of shots on goal, possession information, corners, off sides, fouls, cards given, injuries, substitutions, etc. There is scope for the collection of larger data sets, such as the position-per-time of the ball, and each player on the field throughout the game, or on a short video clip. From this complex data set,

David C. Wyld et al. (Eds) : IPPR, SOENG, DaMi, CSIT, AIS, CSE, CSIP, CCNET - 2018
pp. 119–129, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80509

the objective is to detect specific and semantically meaningful events like player ball possession, team ball possession, kick or shoot, etc. Researchers from all over the world have been working for more than a decade to find different solutions for the video analysis. Their research in the domain of event processing is more focused on structured data. However, there are several applications for event driven systems based on image data. Therefore, there is a need for a system that can process multimedia events [1] from images and videos.

In this paper, the proposed framework attempts to detect different events. Images are given as an input to the object detector \Single Shot Multi-Box Detector" (SSD), which provides us with objects expressed in terms of bounding boxes with a given confidence score. We will use this system as a filter because the objects associated with confidence score higher than a specific threshold will be the input to the event detection system for detecting events. Then based on the distance between the bounding boxes of objects and using logical and temporal operators, events are defined.

2. RELATED WORK

Until the discovery of deep learning, sports video analysis, especially soccer video analysis, has been classified into two categories: object tracking and pattern recognition [21, 9]. The use of customized cameras [14] results in computational cost in case of object tracking, whereas the pattern recognition methodology simply extracts lower-level features and then uses a classifier to detect higher level events. A few methodologies which have been used with noticeable success for soccer activity recognition include: Qian et al [17] categorization of events into distinct categories like shoot, goal, etc. Such an approach includes feature extraction and heuristic rules for detecting events. They perform low-level analysis to detect marks (field, lines, logo, arcs, and goalmouth), player positions, ball position, etc, and then derive mid-level features using these cues. In the end, they developed a rule-based system to detect salient events like the goal, corner, etc. Jin et al [10] applied a Hidden Markov-based algorithm for video event detection based on cues fusion and integration. Detecting higher-level events from lower-level events is an important and challenging problem for soccer video analysis. The detection reveals, e.g., the movement of the players and the ball on the field, which could be used to identify certain actions ('passing the ball', 'shot on goal', etc.) or to better understand the overall trend of the game.

Since 2012, deep learning methods such as Convolutional Neural Networks and Restricted Boltzmann Machines have been successfully used for event and activity recognition. CNNs have shown better performance in image classification, object detection and modeling high-level visual semantics [11],[8],[6]; Recurrent Neural Networks have shown good results in modeling temporal dynamics in videos [12]. Frequently used action localization techniques, such as fast r-CNNs and faster r-CNNs [18],[7], usually start with the region of interest (proposal generation) to obtain a set of candidate regions, then use a fully connected layer at the end to classify objects.

Current approaches mentioned above focus on event recognition in soccer videos from the perspective of feature extraction, models, and classifiers for extracting low-level events. Such approaches lack the semantically meaningful representation of intermediate events. Injecting semantic definition and structural knowledge in these approaches is rather difficult. So, this motivates us to start from the basic building blocks and rebuild a system that allows exploiting the semantic knowledge about events, which can be used to recognize the intermediate and high-level complex events. To the best of our knowledge, while there are systems that automatically

detect basic facts, like the position and the movement of the player, there are no automatic detectors for semantically complex events, like scoring on a penalty kick, or scoring on a corner kick.

The rest of the paper is organized as follows. Section 3 describes the video events as simple and complex events. Section 4 elaborates distinct types of events for the soccer scenario. In Section 5 the proposed architecture is highlighted and in Section 6 results and future work are presented, respectively. In section 7 we draw some conclusions.

3. VIDEO EVENTS

A precise ontological definition of event is still an open point. To the purpose of this paper we take the approach recently proposed in [4]. The main objective of this section is to precisely define the event structure we will adopt in our approach.

Video events can be defined as interesting events which capture the user attention [20] . For example, a soccer "shot on goal" event is defined as the ball kicked by a player and the ball moving towards the direction of the goal.

3.1 Simple Events

A simple event type is defined as follows:

$$SE = \langle ID, seType, t, \langle role_1, oType_1 \rangle, \ldots, \langle role_n, oType_n \rangle \rangle \quad (1)$$

where *ID* is the identifier, *seType* is the event type, e.g. "*throwing the ball*", and *t* is the time instant in which the event occurs, $role_1$..., $role_n$ (n = 1,....,n_{max}) are the roles that different objects play in an event of this type, e.g. one role of simple event "*throwing the ball*" is the subject who throws and a second role is the thrown object; finally o$Type_i$ is the legal type of object that can play the role $role_i$, e.g., it is only players who can throw, and only balls can be thrown. Summing up, the complete definition of the event type "*throwing the ball*" is

$$\langle ID, Throwing\_the\_ball, t, \langle throwing\_Player, player \rangle, \langle throwed\_Object, ball \rangle \rangle$$

A specific instance of an event of simple type defined in (1) is the following tuple:

$$\langle ID, seType, t, \langle role_1, O_1 \rangle, \ldots, \langle role_n, O_n \rangle \rangle$$

where *ID* is the event identifier, O_1 and O_n are identifiers of objects detected in the frame associated to the time *t*, respectively. The instance of "Throwing_the_ball"

$$\langle 12, Throwing\_the\_ball, t, \langle throwing\_Player, obj02 \rangle, \langle throwed\_Object, obj01 \rangle \rangle$$

describes a simple event of type "throwing the ball" that happened at time t, where the obj02 throws the obj01. Furthermore obj01 and obj02 are two objects detected in the frame corresponding to time t, of type ball and player respectively.

3.2 Complex Events

Complex events are built by appropriately aggregating events, previously defined. More precisely, starting from simple events, we can apply logical operators or temporal operators to build higher-level complex events. We can thus define the hierarchy of events, from the lowest level including the simple events to the higher and higher levels corresponding to more and more complex events. In the following, we define the two categories of complex events: logical complex events and temporal complex events.

- **Logical Complex events** A logical complex event stems from the application of logical operators like AND, OR, NOT to a set of events which may be simple or complex.

$$LCE = \langle ID, ceType, t, L =< e_1\ op\ e_2\ op....op\ e_n > \rangle$$

  where *ID* is the event identity, *ceType* is the complex event type (such as "The goal is valid only if there is no foul"), *t* is the time instance in which the complex event occurs, *L* is the set of lower-level simple or complex events e_1. . . . . e_n joined by logical operators *op* (*i.e. AND, OR, NOT*).

- **Temporal Complex events** A temporal complex event derives from the application of temporal operation THEN as follows:

$$TCE = \langle ID, ceType, t, L =< e_1\ THEN\ e_2...THEN\ e_n > \rangle$$

  where *ID* is the event identifier, *ceType* is the complex event type (such as “player 1 passes the ball to player 2"), *t* is the event occurrence time, *L* is the sequence of lower-level simple or complex events, e_1. . . . . e_n that must occur in the order. For example, e_1, e_2, e_3, e_4 may be, respectively, “player1 possesses the ball", “player1 kicks the ball", “the ball approaches player 2", “player2 gets in possession of the ball".

4. TYPES OF EVENTS

One of the most interesting things about soccer analysis is the ability to recognize events, such as a kick, goal, pass, offside, cards, ball possession, etc. from a common video. Most of the videos previously used in the event recognition use multiple fixed cameras to observe the position of all the players and the ball on the soccer field [5]. The use of such cameras improves the overall accuracy of the system for object tracking but they are computationally expensive. The fragment of video we have used can be easily accessible from the internet.

In this section, we try to define a few of the significant low or intermediate complex events in soccer video (consisting of a sequence of frames), such as ball possession and kicking the ball based on the distance between the bounding boxes of involved objects, and rules (combination of temporal and logical operators) defined for each event category.

In this first attempt we propose a rule-based definition of video events, but we are aware that this will turn out to be not very flexible, and in the future we will investigate on the possibility of automatically learning event detectors by using supervised machine (deep) learning techniques.

4.1 Ball possession Event

Ball possession can be classified as Player Ball Possession (PBP) and Team Ball Possession (TBP). Both have the same starting point but different end-points [13]. In our approach, only those time intervals in which the ball is in play are considered for determining the ball possession. When the ball is in play one of the two teams always has the ball possession. PBP starts immediately as soon as a player begins to perform an action with the ball and ends when the player is no more able to perform any action with the ball or there is game interruption.

Player ball possession can be formally defined as follows: the event occurs when the distance between a player and the ball is below a threshold value and that player is the nearest to the ball.

$$\Big\langle ID, PlayerBallPossession, t+\tilde{k}, \langle PossPlayer, p_i\rangle, \langle PossObject, b\rangle \Big\rangle \leftarrow$$
$$player(p_i), ball(b), D(p_i, b, t) < T_h,$$
$$\forall j \neq i, player(p_j), D(p_j, b, t) > D(p_i, b, t) \wedge$$
$$\forall k = 1 \ldots \tilde{k}, D(p_i, b, t+k) \approx 0$$

The event "Player Ball Possession" occurs at time $t+\tilde{k}$, when the distance D (p_i,b,t) between the player p_i and the ball b at time t is less than the threshold T_h, and the distance D (p_j,b,t) between the ball and any other player p_j, $j \neq i$, is greater than D (p_i,b,t). Also, after interaction, the distance between the player and the ball is very low for an appropriate number of $\tilde{k}$ consecutive frames. The value T_h determines the threshold value for a player being able to physically interact with the ball and must be calculated experimentally.

4.2 Kicking the ball Event

In the soccer video, with reference to the consecutive sequence of frames, the event corresponding to kicking the ball is identified, initially if the distance between a player and the ball is very low for a few frames. Then, if the distance between a player and the ball increases in an appropriate number of the subsequent frames and the player is no longer able to interact with the ball. We can formally define the event Kicking the ball as follows:

$$\Big\langle ID, KickingTheBall, t+\tilde{k}, \langle KickingPlayer, p_i\rangle, \langle KickedObject, b\rangle \Big\rangle \leftarrow$$
$$player(p_i), ball(b), D(p_i, b, t) < T_h \quad \wedge$$
$$\forall k = 0 \ldots \tilde{k}-1, D(p_i, b, t+k) < D(p_i, b, t+k+1)$$

The expression above holds true as long as the distance between the player and the ball increases after their interaction. T_h is the interaction threshold between the player and the ball. In a game Kick can be classified into several types: Free kick, Goal kick, Penalty kick, Corner Kick etc.

4.3 Limitations

While defining the events we are not considering all special cases that might occur during a match. In some cases, the player does not interact with the ball, and runs besides the ball without touching it. Player ball possession only starts with the first touch. Also, considering ball possession for the player nearest to the ball is wrong, e.g, when that player is standing with back to the ball. To better differentiate between kick or shoot and dribble, one can think of the speed with which the ball travels after the player ball interaction. For example, the speed of the ball after dribbling will be slower than that of kicking or shooting. We are also considering the same threshold for all the players as taking into account player profiles related to their typical interaction with the ball is out of the scope of this work.

5. PROPOSED ARCHITECTURE

Figure 1 describes the workflow for our methodology. The data at our disposal consist of approximately 5 mins long video, consisting of 7.5k annotated frames. Objects are detected from every single frame using SSD [15]. Then a specific threshold regarding the confidence score is defined to filter out the objects which are not required to define events. Finally, events will be detected based on the distance between the bounding boxes of objects using temporal and logical operators.

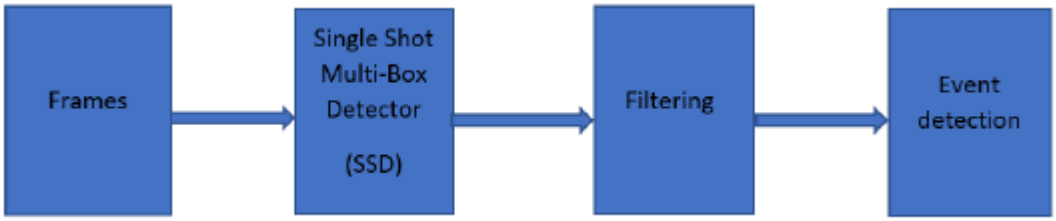

Fig.1. Block diagram of the proposed architecture

Frame Data We have a sequence of frames { $f_1, f_2 ..... f_n$ } Each frame is a set of bounding boxes, each bounding box gives us the position and dimension of an object, such as the ball or a player, by specifying the coordinates of the region containing the object. Frames are given as input to the SSD to detect objects with a confidence score.

Single Shot Multi-Box Detector (SSD) Most of the methods previously used for object detection have one thing in common, they have one part of their system dedicated to providing region proposals which includes re-sampling of pixels and features for each bounding box, followed by a classifier to classify those proposals. These methods are useful but are computationally expensive resulting in low frame rate. Another simpler way of doing object detection is by using a high-speed SSD system, which combines the two tasks of region proposal and classification in one system. The key idea behind SSD is small convolutional filters are applied to feature maps of bounding boxes to predict the category scores, using separate predictors for different aspect ratios to perform detection on multiple scales.

SSD needs an input image and ground truth for each object class during training. We have created this training set starting from a fragment of a real soccer match video, using Vatic [22], a Video Annotation Tool. Vatic allows annotating objects inside each frame drawing a bounding box

around them. The output of this process is a set of images with relative bounding boxes coordinates saved in PascalVOC format.

Table 1 shows the numbers of object manually annotated, used for the training and test of SSD.

Table 1. Objects manually annotated to train and test the SSD

	Ball	Player	Goal	Player Name	Flag
Training	1839	22756	534	542	887
Test	593	4725	134	208	223
Total	2432	27490	668	750	1110

The training set in Table 1 has been used to create the SSD model. The average precision on the test set is given below in Table 2.

An example of the input image and the output image from the soccer match to SSD is shown in Figure 2 and Figure 3, respectively.

Filtering Filtering is performed by defining a specific threshold for the objects detected by the SSD. For example, as multiple players are detected in a single frame, then using a specific threshold, we can discard players in the frames which are not necessary to define the action.

Table 2. Average precision of the system

Ball	0.776696
Player	0.904298
Flag	1.0
Goal	0.999327
Player Name	0.907692

Event Detection System In many application domains, such as video event activity detection, sequences of events occurring over time need to be studied to summarize the key events from the video clips [19]. This section deals with the specific strategies adopted by the system for event detection. The steps involved are the detection and collection of the simple and low-level complex events, and the composition of the same to detect higher-level complex events. The system also includes an event type to identify the class of events. The new incoming event is registered within the system with a unique event identifier. The event recognition is performed by means of monitoring routines at two levels, low-level recognition and high-level recognition. The low-level event recognition involves detection of simple primitive events, while high-level event recognition handles detection of complex events. An event detection system receives, as an input, bounding boxes associated with a confidence score. Each bounding box also represents the coordinates of the object. To recognize the higher-level complex event, the system first detects simple and low-level complex events based on the rules defined for each event category and stores those events in the memory. We then apply logical and temporal operators on the detected events to recognize the higher-level complex events. Although there are several programming languages available to implement the event detection system, python was our preferred choice because of its highly intuitive general-purpose syntax.

Fig. 2. Original frame

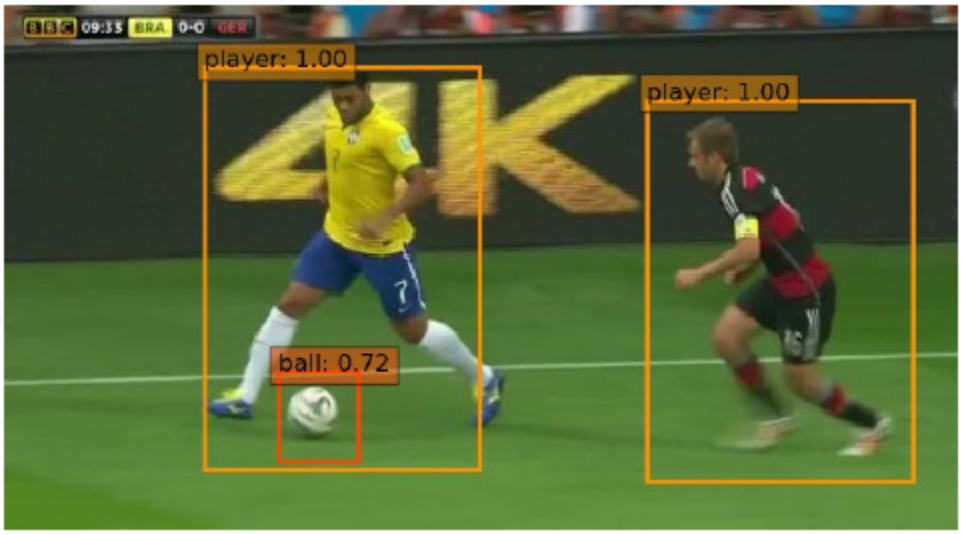

Fig. 3. Objects detected by SSD with confidence score

6. RESULTS AND FUTURE WORK

We have applied the proposed system to detect low-level complex events like "ball possession" and "kicking the ball" in the real soccer video. We have experimented on 5 minutes short video consisting of approximately 7.5k frames. We are aware of the fact that a limited number of events can be detected from this small data set. In the future, we will experiment on a larger data set, thus the number of events can be increased. Table 3 shows the event detection results. For Ball possession event, 13 out of 14 events have been detected successfully, one event was missed as in few frames two players are very close to each other, so it is hard to recognize possession. In our

experiments, the detection of such events occurs if the event definition is met for an appropriate number of consecutive frames. In this very preliminary application of the proposed event detection framework, we referred to a heuristically chosen number of consecutive frames equal to 5. For example, if the distance between the ball and the player is very low for five consecutive frames, we have a Ball possession event.

Table 3. Event detection results

Detected Events	Total	detected	Miss	Accuracy
Ball Possession	14	13	1	92%
Kicking the ball	19	16	3	84%

In the next consecutive sequence of frames, if the distance between the ball and the player increases with respect to a specific threshold in an abrupt manner, we have a kicking the ball event. For kicking the ball event, 16 out of 19 events were detected successfully, three events were missed as in some cases it may happen that when the players kick the ball, the ball hits the next closest player in fewer than five frames.

In the future, based on the simple and low-level complex events, we are planning to detect more complex events such as "Pass the ball" and \Shot on goal" by effectively merging the simple and low-level complex events using logical and temporal operators. To define the higher-level complex events, we have taken into consideration events at different abstraction levels. To define the event "Pass the ball" let us consider Player1 and Player2 of the same team. While referring to players of the same team let us assume that the color of the upper half of the bounding box is the same. For instance, the higher-level complex event "Pass the ball" basically occurs if the following lower-level complex events occur. With respect to the successive sequence of frames, the event corresponding to "Player1 is in possession of the ball" is identified, if the distance between Player1 and the ball is very low for a few frames. Then, if the distance between Player1 and the ball increases in an appropriate number of the subsequent frames we can define the low-level complex event as "Kicking the ball". In the same consecutive sequence of frames if the distance between Player2 (of the same team as Player1) and the ball decreases up to a very low value and the possession of the ball is with Player2, while there is no other object between the ball and Player2, then we can define the higher-level complex event as "Pass the ball":

$$\left\langle 23, Pass, t+\tilde{k}, \langle passingPlayer, p_1\rangle, \langle receivingPlayer, p_2\rangle, \langle passedObject, ball\rangle \right\rangle$$

where 23 is the identifier, Pass is the event type, $t+\tilde{k}$ is time instance in which the event occurs. *passingPlayer* is the role performed by p1 on object ball, *receivingPlayer* is the role performed by p2 on object ball.

To define the event "Shot on goal" let us consider the three entities player, ball and goal post. The higher-level event “Shot on goal" basically occurs if, with reference to the consecutive sequence of frames, the player kicks the ball, the distance between the ball and the player increases and the distance between the ball and the goal post decreases up to a specific threshold. Then we can define the higher-level event as "Shot on goal":

$$\left\langle 20, ShotOnGoal, t+\tilde{k}, \langle KickingPlayer, p\rangle, \langle KickedObject, ball\rangle, \langle GoalPost, G\rangle \right\rangle$$

where 20 is the identifier, *ShotOnGoal* is the event type, $t+\tilde{k}$ is the event occurring instance, *KickingPlayer* is the role performed by *p*, *GoalPost* is the role of object *G*, when object ball approaches towards it.

7. CONCLUSIONS

In this paper, we have defined a few simple and complex events for the soccer video. We have also proposed a distance-based event detection system. The event detection system takes as an input bounding boxes associated with a confidence score for each object category. The system successfully detects the low-level complex events, such as: "Ball possession" and "Kicking the ball ". The results demonstrate the validity and the effectiveness of our methodology.

REFERENCES

[1] Challenges with image event processing, 2017. Poster DEBS 17.

[2] Adnan Akbar, Francois Carrez, Klaus Moessner, and Ahmed Zoha. Predicting complex events for pro-active iot applications. In Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on, pages 327{332. IEEE, 2015.

[3] Adnan Akbar, Abdullah Khan, Francois Carrez, and Klaus Moessner. Predictive analytics for complex iot data streams. IEEE Internet of Things, 2017.

[4] Stefano Borgo and Riichiro Mizoguchi. A first-order formalization of event, object, process and role in yamato. In FOIS, pages 79-92, 2014.

[5] Pascual J Figueroa, Neucimar J Leite, and Ricardo ML Barros. Tracking soccer players aiming their kinematical motion analysis. Computer Vision and Image Understanding, 101(2):122-135, 2006

[6] Sebastian Gerke, Karsten Muller, and Ralf Schafer. Soccer jersey number recognition using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 17-24, 2015.

[7] Ross Girshick. Fast r-cnn. arXiv preprint arXiv:1504.08083, 2015.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580-587, 2014

[9] 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.

[10] Guoying Jin, Linmi Tao, and Guangyou Xu. Hidden markov model based events detection in soccer video. Image Analysis and Recognition, pages 605-612, 2004

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097-1105, 2012.

[12] Guang Li, Shubo Ma, and Yahong Han. Summarization-based video caption via deep neural networks. In Proceedings of the 23rd ACM international conference on Multimedia, pages 1191-1194. ACM, 2015.

[13] Daniel Link and Martin Hoernig. Individual ball possession in soccer. PloS one, 12(7):e0179953, 2017.

[14] Jia Liu, Xiaofeng Tong, Wenlong Li, Tao Wang, Yimin Zhang, and Hongqi Wang. Automatic player detection, labeling and tracking in broadcast soccer video. Pattern Recognition Letters, 30(2):103-113, 2009.

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21-37. Springer, 2016.

[16] L. Pappalardo and P. Cintia. Quantifying the relation between performance and success in soccer. ArXiv e-prints, May 2017.

[17] Xueming Qian, Guizhong Liu, Huan Wang, Zhi Li, and Zhe Wang. Soccer video event detection by fusing middle level visual semantics of an event clip. In Pacific-Rim Conference on Multimedia, pages 439-451. Springer, 2010

[18] 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, 39(6):1137-1149, 2017.

[19] Wei Song and Hani Hagras. A big-bang big-crunch type-2 fuzzy logic based system for soccer video scene classification. In Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on, pages 2059{2066. IEEE, 2016.]

[20] P Thirumurugan and S Hasan Hussain. Event detection in videos using data mining techniques. International Journal of Computer Science and Information Technologies, 3(2):3473-3475, 2012.

[21] Dian W Tjondronegoro and Yi-Ping Phoebe Chen. Knowledge-discounted event detection in sports video. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(5):1009-1024, 2010.

[22] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowd sourced video annotation. International Journal of Computer Vision, 101(1):184-204, 2013.