Soccer Event Detection

Michele Attilio Iodice
matricola: 09657A
Università degli studi di Milano

February 24, 2024

1 Introduction

Football is one of the most famous and followed sports in the world; its allure attracts millions of spectators. Thanks to the giant strides made by technology in recent years, those working in this sector have been able to leverage technical support for analysis and statistics.
Several studies have delved into this field, developing algorithms and systems capable of providing useful statistics to be used as information for spectators or as support for teams to evaluate their performances. For instance, estimating team tactics[1], tracking players[2] or the ball on the field[3], detecting events occurring in the match[4] summarizing the soccer match[5], and estimating ball possession statistics[6].
These systems have been developed through the study of new techniques and methods, such as machine learning (ML), for image and/or video processing. Artificial intelligence (AI), particularly machine learning, comes to aid in this context, providing intelligent and much more precise results.
Machine learning algorithms can be developed in various ways, their purpose in this context is to detect football-related events and generate useful statistics for those working in this field.
Event detection is a process that aids in obtaining statistics, and distinguishing between the different types of events that can occur during a match simplifies the operations to create a sort of synthesis of a football match.
The purpose of this study is to detect events during a match. To do this, various machine learning architectures have been developed. To train these architectures, it was necessary to use a dataset representing the events we want to recognize.
For the study, the SEV dataset[7] was used, within which images of 7 main match events are collected: penalty kick, corner kick, free kick, tackle, substitution, yellow card, and red card, along with the three playing areas: center, right area, left area. And a set of images that are not considered important. This dataset is then used to train our network.
First, the network must be able to distinguish between images that belong to the match and those that do not, as it may happen that other images are transmitted during the match, such as advertisements, the audience, or close-ups of players that are irrelevant for event detection purposes. Therefore, a first VAE (variational autoencoder) model was developed to distinguish whether an image is a highlight or not.
Once it is recognized as an image related to the match, a network that classifies the images detects which event among those mentioned earlier it corresponds to.
Experiments show that using the VAE before the classifier improves the classifier's performance from 18% to 40% accuracy. The performance further improves with the use of a new fine-grain classification network to classify yellow and red cards.
In the rest of the report, in section 2, we will provide insights into the three developed networks and their architectures. In section 3, the dataset used for each network, the proposed algorithm, how the trainings were conducted, the metrics and their results for each model, and how tests were conducted on the entire system will be explained in detail, reporting all the results. In section 4, some conclusions will be drawn, and some future work will be proposed.

1.1 Related Work

One of the first work in this area is the study of Sigari et. al.[8] that employ the fuzzy inference system. The algorithm presented in this method works based on replay detection, logo detection, view type recognition and audience excitement. This method is also limited to three events: penalty, corner, and free-kick
The work of Lin et. al.[9] is one of the researches in the field of fine-grained image classification, which is based on deep learning. In this model, two neural networks are used simultaneously. The outer product of the outputs of these two networks is then mapped to a bilinear vector. Finally, there is a softmax layer to specify the classification of images. The accuracy of this method for the CUB-200-2011 dataset is 84.1%
In 2018, Sun et. al.[10] presented new architecture on an attention-based CNN that learns multiple attention region features per an image throug the one-squeeze multi-excitation (OSME) module and then use the multi-attention multi-class constraint (MAMC). Thanks to this structure, this method improves the accuracy of previous methods to some extent.
The Soccer Event Detection Using Deep Learning research of Karimi, A., Toosi, R., Akhaee, M. A.[11] is one of the last study in this area disclosed in 2021. In the paper, is proposed a deep learning approach to detect events in a soccer match emphasizing the distinction between images of red and yellow cards and the correct detection of the images of selected events from other images. The method includes three modules: the variational autoencoder (VAE) module to differentiate between soccer images and others image, the image classification module to classify the images of events, and the fine-grain image classification module to classify the images of red and yellow cards. They introduce a new dataset for soccer images classification that is employed to train the networks mentioned in the paper.

2 Research question and methodology

What we aim to achieve with our method is for it to accurately distinguish whether an image is a highlight or not, and if it is a highlight, associate it with the correct corresponding event.
In this section, we delve into the heart of our system by providing details on the proposed method, closely examining all three components that comprise it, which include the image classifier to detect the event associated with the image, a fine-grained classification module to distinguish between yellow or red cards, and the variational autoencoder to detect whether it is an highlight or not.

2.1 Proposed Method

The proposed method is divided into three parts. First, the image is passed through the variational autoencoder. If the loss of the VAE network is less than a certain threshold, then the input image is considered to be a match event and passed to the image classifier.
The image classification module assigns a specific event to the input image. If the event corresponds to one of these categories: center circle, right penalty area, or left penalty area, then they are classified as no highlights. Otherwise, it is classified as one of the following events: penalty kick, corner kick, tackle, free kick, or substitution. If the classified event falls into the Cards category, the image is then passed to the fine-grain classification module, which distinguishes whether the image is a red or yellow card, associating the card's color with the image.
The method just described is depicted in Fig 1. In the following paragraphs, we will describe in detail the three modules.

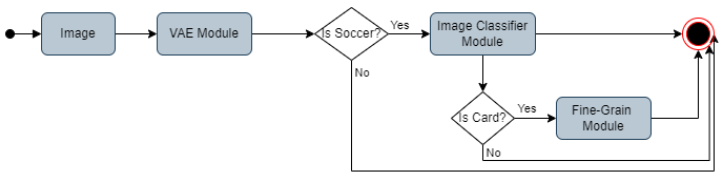


Figure 1: Diagram of the proposed algorithm

2.2 Variational Autoencoder Module VAE

During a match, there can be various images that are not always related to the game, such as during commercial breaks, or when the camera focuses on fans in the stands or on the benches, or when there are close-ups of individual players during a break in play. In these cases, these images are categorized as not being part of the football match.To achieve this, there is a need for a module that is capable of recognizing event images compared to other types of images.
The VAE network is therefore used to identify if the input image is similar to those contained in the SEV dataset. Its architecture is shown in Fig 2. The entire training dataset of the SEV dataset[7] is passed to the VAE for training. Then, using the reconstruction loss (obtained as the difference between the input image and the one reconstructed by the variational autoencoder) and determining a threshold value on it, images that have a higher loss value than the threshold are considered as no highlights, while images that have a lower loss value than the threshold are considered to be soccer game images. Therefore, the VAE acts as a two-class classifier, putting them into the no highlight class or soccer game images class.

Figure 2: Variational autoencoder module architecture (VAE)

2.3 Image Classification Module

The image classification module is tasked with associating one of nine categories with the input image with a certain degree of accuracy. If this accuracy value is below a certain precision threshold, the image is considered a no highlights image.
This threshold is set to a very high value specifically to avoid classifying images that do not belong to the football match. Additionally, if the image is classified as belonging to one of the following classes: left penalty area, right penalty area, and center circle, then it is also considered a no highlights image.
However, if it falls into one of the following classes: penalty kick, corner kick, tackle, free kick, and substitution, the event is associated with the image and reported as the output of the network; whereas if it is classified as a card, then the image is passed to the fine-grain module.
The architecture of the network is based on the pre-trained EfficientNetB0 model with ImageNet weights, to which final dense layers are added, as shown in Fig 3.

Figure 3: Image classification module architecture

2.4 Fine-Grain Classification Module

When the card category is associated with the image, it is necessary to distinguish the type of card, whether it is red or yellow. The images in these two categories are exactly the same; the only difference lies in the color of the small card. Therefore, if we had classified them together with the other classes, it would have been very difficult to distinguish between the two cases, and there would have been many errors. For this reason, a specialized network was developed solely to distinguish images representing the card and separate them into two classes: red card and yellow card. The network is based on the pre-trained EfficientNetB0 model with ImageNet weights. The network is divided into two branches composed of different processing layers that are combined through tensor product, as shown in the architecture depicted in Fig 4.

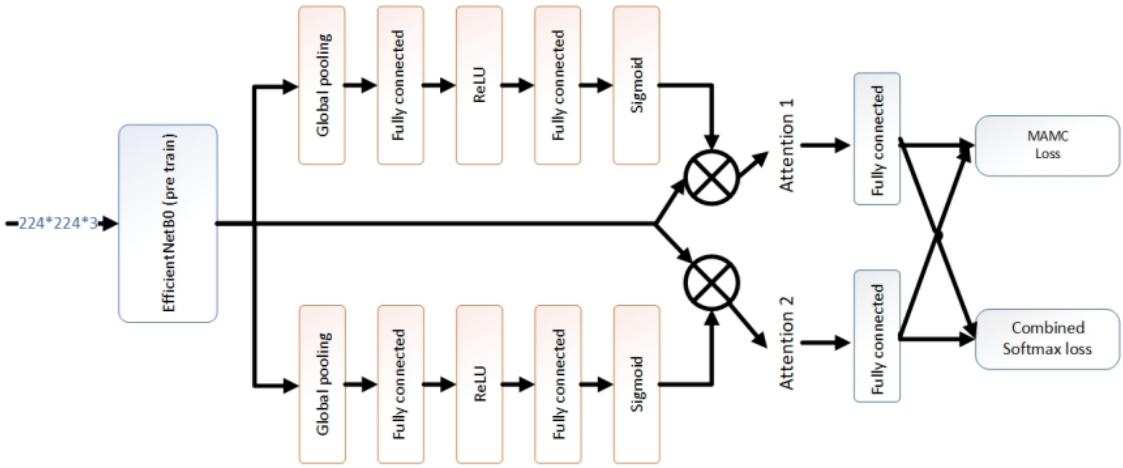


Figure 4: Fine-grain classification module architecture

3 Experimental results

3.1 Dataset

In this study we use a Soccer Event Dataset (Image) were was collected two image datasets of a football match:

- the Soccer Event (SEV) dataset covering the football match events.
- the Test Event dataset used to assess the proposed architecture.

The aforementioned datasets were collected in two ways:

- Web crawling and collection of images related to events.
- Watching the videos of UCL and European league football matches to extract the relevant frames.

The SEV dataset covered 7 football match events and 3 scenes from the football field. The seven main events include: **Corner Kick, Penalty, Kick, Free Kick, Red Card, Yellow Card, Tackle, substitute** as show in Fig5a. The scenes from the football field included: **Left Penalty Area, Right Penalty Area, Center Circle** as show in Fig5b. The Soccer event dataset consists of 60000 images from 10 different events, each of which includes 6000 images.In each category, 5000 images are used as the training data, 500 images are used as validation data and 500 images are used as test data. The Test Event dataset includes 4,200 images falling within 3 classes as show in Table1.

(a) (b)

Figure 5: Samples of soccer event dataset(left) and Samples of test event dataset(right)

Class Name	**Image Number**
Soccer events	1400
Other soccer events	1400
Other images	1400
Sum	4200

Table 1: Test Event Dataset

3.2 Training

The three networks were trained separately on the training dataset mentioned in the previous section. The training methods of the three networks, namely VAE, image classification, and fine-grain classifier, are detailed in the following subsections.

3.2.1 VAE training

To train the VAE network, the 7 events defined in the SEV dataset are used as input data for training. These data are also utilized for testing and evaluating the model's performance. The training specifications are outlined in the table2. As evident from the figure6, the loss curve value decreases with the number of epochs. So the the value of 144.5 is choice as the threshold for the loss of the method, becouse it gives the best distinction between categories.

Parameter	**Value**
Optimizer	Adam
Loss function	Reconstruction loss + KL loss
Preformance metric	Loss

Table 2: Simulation parameters of the variational autoencoder

3.2.2 Image Classification training

The image classifier was first trained on the pre-trained EfficientNetB0 model on the ImageNet collection with dimensions 224 * 224 * 3. Subsequently, using transfer learning, the network was trained on the SEV dataset with input images of dimensions 224 * 224 * 3. The network underwent 20 epochs using the parameters defined in the table3. Regarding the images of the yellow card and red card classes, they were combined into a single class "Cards" so the training classes are 9. The figure7 shows the trend of

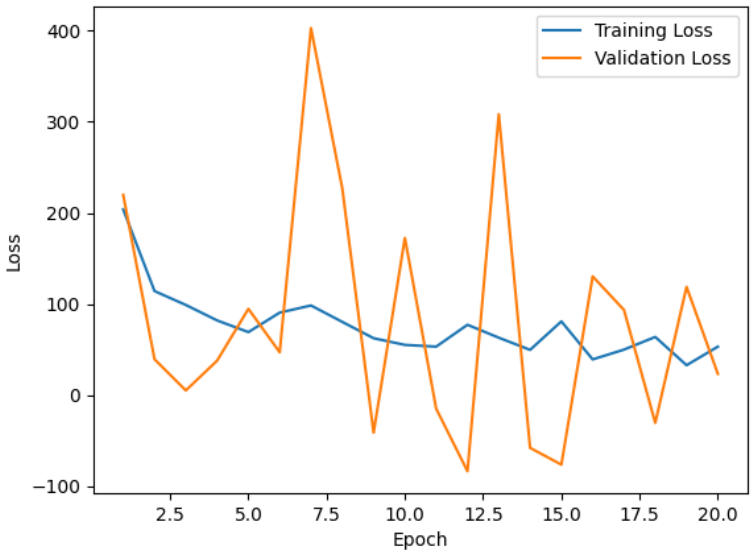


Figure 6: Loss curve of VAE module

the loss values for training and evaluation of the model over the epochs. The accuracy of the proposed method for image classification is 18.11%.

Parameter	**Value**
Optimizer	Adam
Loss function	Categorical Cross-Entropy
Preformance metric	Accuracy
Total Classes	9 (red and yellow card classes merged)
Augemnation	Scale, Rotate, Shift, Flip
Batch Size	16

Table 3: Simulation parameters of the image classification module

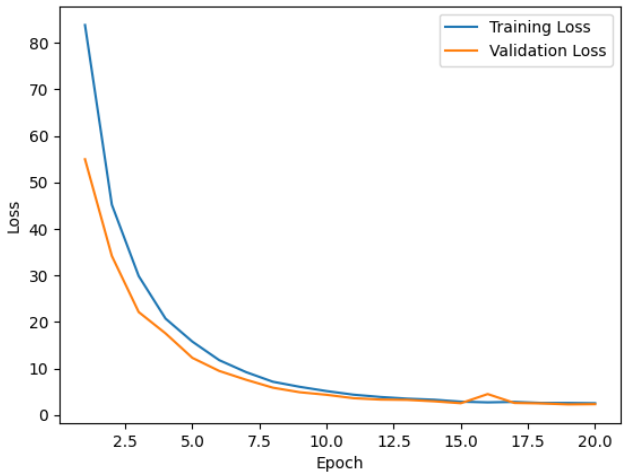


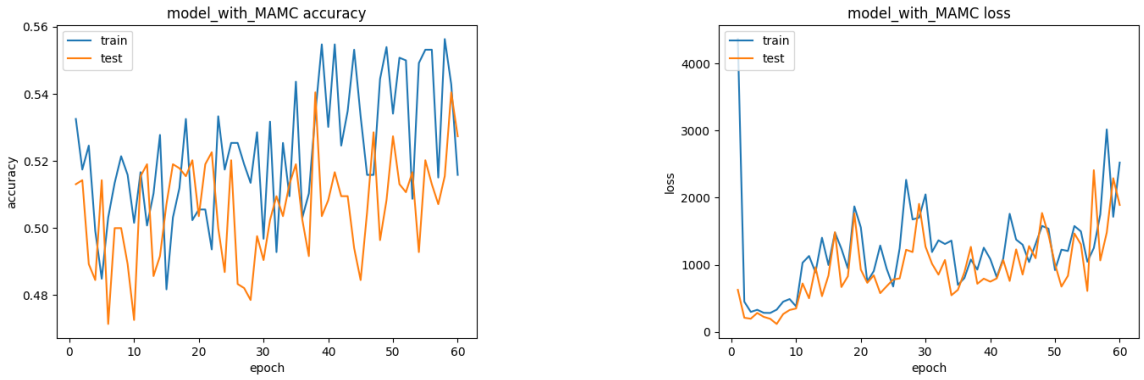
Figure 7: Loss curve of image classification module

3.2.3 Fine-grain classification training

The fine-grain classification network was trained on two classes of the SEV dataset: the red-cards and yellow-cards classes. For each category, 5000 images were used for training, and 500 images were used for both evaluation and testing. The parameter specifications used for this training are listed in the table5. Figures 8a and 8b respectively show the loss and accuracy values across epochs.The introduction of this distinction between the cards, return an accuracy in the train data of 53.6% of the image classification for the two classes. The reasons for this uncertainty are due to the fact that the two classes only differ in a single detail (the color of the cards).

Parameter	**Value**
Optimizer	Adam
Loss function	Categorical cross-entropy
Preformance metric	Accuracy
Total Classes	2 (red and yellow card classes)
Augemnation	Scale, rotate, shift, flip
Batch Size	16

Table 4: Simulation parameters of the fine-grain image classification module



(a) (b)

Figure 8: Trend of accuracy on the left and trend of losses on the right over the epochs.

3.3 Evaluation Metrics

Different metrics are exploited to evaluate this network. In order to evaluate the image classification architectures and Fine-grain module networks. The accuracy metric is used as the main metric; recall and F1-score are also used to determine the appropriate threshold value of the network. Also, precision is used to evaluate the performance of the proposed method to detect events.
The accuracy metric can be used to determine how accurately the trained model predict in the image classification module and fine-grain classification.

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

The F1 score metric considers both the recall and precision criteria together; the value of this criterion is one at the best-case scenario and zero at the worst-case scenario.

$$F_1 = \frac{2}{Precision^{-1} + Recall^{-1}} = \frac{TP}{TP + \frac{1}{2} \times (FP + FN)}$$

Precision is a metric that helps to determine how accurate the model is when making a prediction. This metric has been used as a criterion in selecting the appropriate threshold.

$$Precision = \frac{TP}{TP + FP}$$

Recall metric refers to the percentage of total predictions that are correctly categorized .

$$Precision = \frac{TP}{TP + FN}$$

3.4 Test and Evaluation of the Proposed Algorithm

The various components of the proposed method were evaluated separately to find the best model for classifying the images.
The three trained models were then combined into a single algorithm in which the images are first passed to the VAE model.If the model returns a loss value lower than the threshold set at 144.5, then it is an image of the football match, which is then passed to the image classification model. If this model returns an accuracy value equal to or greater than 0.1, then it is considered an event, and the corresponding class is associated with it. The use of this low value is caused by the low accuracy of the model, so for test the algorithm we use this value, Otherwise, no class would be returned. (In future we re-train the model for improve the performance and set the threshold at 0.9). If it falls into the Cards class, the image is finally passed to the fine-grain classification model to distinguish between a red or yellow card, otherwise the corresponding class is returned, unless it falls into one of the following classes:**center, right area, or left area**. This algorithm was tested using the Test Event dataset described in section 3.1. The evaluation results are shown in the table.

Class	Sub-class	**Precision**
Soccer Events	Corner-Kick	0.11
	Free-Kick	0.12
	Penalty-Kick	0.11
	Red-Cards	0.95
	Yellow-Cards	0.99
	Tackle	0.12
	To-Subtitute	0.11
Other-soccer-events		0.11
Other-images		0.99

Table 5: The precision of the proposed algorithm

4 Concluding remarks and Future Work

In this report, we discussed a method for classifying images of a football match into events related to the match itself. To assess the reliability of the model, we tested it using the SEV dataset. The tests are divided into three categories: the first includes images that do not belong to the football match, the second includes the 7 events described earlier, and the last includes other images of the match that do not fall into the previous categories. Thus, three methods were proposed, one for each category: the VAE recognizes whether an image qualifies as a highlight, the image classification seeks to determine which event the image corresponds to, and the fine-grain image classification module was used to differentiate between red and yellow cards. By combining these methods, the final model achieves an accuracy of approximately 40.1%, thanks to the use of all three components of the method, which provide better performance than they would individually.This low precision is caused by the image classification module having low accuracy, stemming from the model being trained on a dataset that is too small for the problem. It is not yet able to distinguish between classes accurately, so it should be retrained on a much larger dataset.
Some future developments of this study could include:

- expanding the training dataset by providing more samples for each category.
- introducing new categories of events related to the match.
- developing algorithms for analyzing the recognized events.
- using this method as a data preprocessor for other systems.

References

[1] T. Ogawa G. Suzuki, S. Takahashi and M. Haseyama. Team tactics estimation in soccer videos based on a deep extreme learning machine and characteristics of the tactics. *IEEE Access, vol. 7, pp. 153, 238–153 248*, 2019.

[2] H. Ebadi M. Manafifard and H. A. Moghaddam. A survey on player tracking in soccer videos. *Computer Vision and Image Understanding, vol. 159, pp. 19–46*, 2017.

[3] A. Keskar P. Kamble and K. Bhurchandi. A deep learning ball tracking system in soccer videos. *Opto-Electronics Review, vol. 27, no. 1, pp.58–69*, 2019.

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

[5] R. Muhammad R. Agyeman and G. S. Choi. Soccer video summarization using deep learning. *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.

[6] A. Chakrabarti S. Sarkar and D. Prasad Mukherjee. Generation of ball possession statistics in soccer using minimum-cost flow network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[7] Ali Karimi, Ramin Toosi, and Mohammad Ali Akhaee. Soccer event detection using deep learning. *arXiv preprint arXiv:2102.04331*, 2021.

[8] H. Soltanian-Zadeh M.H. Sigari and H.-R. Pourreza. Fast highlight detection and scoring for broadcast soccer video summarization using on-demand feature extraction and fuzzy inference. *International Journal of Computer Graphics, vol. 6, no. 1, pp. 13–36*, 2015.

[9] A. RoyChowdhury T.Y. Lin and S. Maji. Bilinear cnn models for fine grained visual recognition. *Proceedings of the IEEE international conference on computer vision, pp. 1449–1457*, 2015.

[10] F. Zhou M. Sun, Y. Yuan and E. Ding. Multi-attention multi-class constraint for fine-grained image recognition. *Proceedings of the European Conference on Computer Vision (ECCV), pp. 805–821*, 2018.

[11] Toosi R. Akhaee M. A. Karimi, A. Soccer event detection using deep learning. *arXiv preprint arXiv:2102.04331*, 2021.