A Novel framework for Fine Grained ActionRecognition in Soccer

by

Ganesh Yaparla, Sriteja Allaparthi, Sai Krishna Munnangi, Garimella Ramamurthy

in

15th International Work-Conference on Artificial Neural Networks

Gran Canaria, Spain.

Report No: IIIT/TR/2019/-1

Centre for Communications
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2019

A Novel framework for Fine Grained Action Recognition in Soccer

Yaparla Ganesh, Allaparthi sri Teja, Sai krishna Munnangi and Garimella Rama Murthy

International Institute of Information Technology,Hyderabad, India
ganesh.yaparla, sriteja.allaparthi, krishna.munnangi
@research.iiit.ac.in, rammurthy@iiit.ac.in

Abstract. Sports analytics have become a topic of interest in the field of Artificial intelligence. With the availability of huge volumes of high level data, significant progress has been made in the domain of action recognition in the past. Though video based action recognition has progressed well using state of the art deep learning techniques, its applications are limited to some higher level actions like throwing, jumping, running etc. There has been some work in fine-grained action recognition technique, such as, identification of type of throws in Basketball, and the type of a player's shots in Tennis. However with larger play field and with many players on field, multi player sports such as Soccer, Rugby, Hockey and etc. pose bigger challenges and remain unexplored. These games in general are live fed through field view cameras or skycams which aren't stationary. For these reasons, we chose to recognize player's actions in the game of Soccer and thereby, explore the capabilities of existing architectures and deep neural networks for these kind of games. Our main contributions are the proposed framework that can automatically recognize actions of players in live football game which will be helpful for text query based video search, for extracting stats in a football game and to generate textual commentary and the Soccer-8k dataset which consists of different action clips in the soccer play.

1 Introduction

The problem of Action Recognition requires a system to identify the subject's action from a series of observations on the subject's movements. Action Recognition can be further classified into Sensor-based [1] and Vision-based [2] Action Recognition problems.

Vision-based Action Recognition problems are the most researched problems, with major share taken by Sports Analytics [1, 3–5]. With volumes of rich data available, many video based action recognition techniques are on the rise, but to the best of our knowledge there is no vision based application that delivers some stats related to the sport being played in real time. This might be due to lack of availability of event or player centric datasets for fine grained action recognition in multi player sports. This might be the reason why sensor

based activity recognition predominantly dominates the area of sport analytics. In multi player sports these sensors are used to track people and objects involved so as to extract player specific statistics (heat-maps, traits) and team level statistics(coordination, formations).

We took a step forward to recognize fine grained actions in soccer. Understanding and recognizing complex events in a soccer match like passing, shooting, heading from videos is a challenging task. This is because of the coverage from different angles by multiple cameras and the existence of multiple players in the region of interest. In this paper, we introduce a framework exploiting the recent developments in the deep learning models for action recognition.

The attempts of earlier works in these directions to classify the player action in futsal matches aren't player centric or event centric [3]. However, such frameworks need in depth labeling of the actions at each player in every frame. Due to lack of such big datasets and the huge manual work for creating the required labels made the the application of the well defined neural network architectures limited in sports.

Many models for fine-grained action recognition in sports have been proposed. These fine-gained action recognition models have been based on Two stream convolutional neural networks [6], Optical flow analysis and spatio-temporal analysis based on 3D CNN [2] and RNN's [7]. In this paper, we are particularly interested in the recognizing different actions like passing, shooting, heading, dribbling etc in the game of soccer, based on spatio-temporal analysis. This requires a special dataset that is soccer specific. With lack of such a standard dataset, we created the Soccer-8k dataset. The samples in the Soccer-8k dataset are RGB, monocular video clips generated from Soccer matches, with dynamic background and contains occlusions. These videos include different moving backgrounds with multiple persons involvedd in the scene of interest making the probability of finding the common features for the same labeled videos very less and potentially covering all the features.

We propose a novel framework for event based action recognition in soccer. The framework takes the video as input and outputs the same video with each individual action tagged and bounded within a box, as represented in Fig. 1. The framework consists of primarily two modules, Event Detector and Action Classifier. The Event Detector module identifies the desired events and generates a video clip containing the actions surrounding the event. This clip is inputted to Action Classifier module. We introduce GAWAC (GAussian Weighted event based Action Classifier) architecture, an integral part of Action Classifier module, which classifies the input clip. Other subsequent modules portrays each classified clip onto the original video by tagging the action and confining it within a bounded box in the input. The framework, and the Soccer-8k dataset are the main contributions of this paper.

The rest of the paper is organized as follows. Section 2 refers to the related work done and Section 3 discusses how the Soccer-8k dataset has been created from the Event Detector module of our framework. Section 4 discusses the architecture and implementation details of GAWAC. Section 5 deals with the

experiments done along with their results on the Soccer-8k dataset respectively. Section 6 discusses the results. Section 7 concludes the paper along with details about the future work.

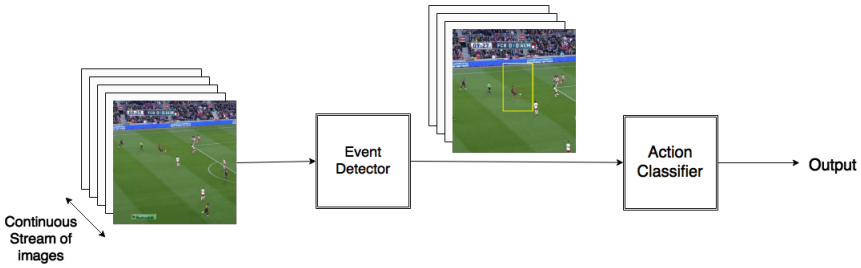

Fig. 1: Block diagram of the framework

2 Related Work

State of the art techniques developed in sport analytics so far are limited to few specific settings. The reason for this is, the problem of Action Recognition incorporates a broad range of scenarios or settings and many other subproblems, each of which can have significant affect. For example, in the case of video based action recognition, an action captured through a dynamic camera needs to be handled differently from the one captured through static cameras. And within the same video itself, monocular view and multi-view may need different processing. With such influential factors and lack of proper standard datasets, the progress made so far is comparatively less.

Many of the previous works using deep learning models in the action recognition deals with the interaction of multiple objects with each other in the scene of interest in different sports. These interactions are used for tracking the ball, finding the possession, person tracking etc. [8–11]. In addition to all of the above models, many have explored the two stream deep learning frameworks for action recognition which deals with temporal and optical flow [6, 7, 12–17]. Along with these deep learning models, some of the existing methods have been exploring the hand-made features like HOG, HOF, MBH extracted from trajectory-based representations computed using optical flow [18–21].

Deep learning models based on 3D CNN and RNN's have shown promising results on fine grained video based action recognition. The reason being 3D CNN and CNN plus LSTM based architectures have capability to extract spatio-temporal information present in the video dataset. [4] have presented a 3 layered LSTM model for fine grained action recognition in the game of Tennis, trained on their custom made dataset, THETIS [22]. First, the authors have extracted the features through the Inception model [23] which were then used to train the

model. [5] provided three LSTM based frameworks for scoring athletic events, within each of which, clip level features are aggregated for video level description differently providing different expressions on quality of the action, that helped in scoring the actions. [3] presents a hierarchical model of CNN and LSTMs for team activity recognition in Soccer. The authors have used 3 classes, Shoot, Pass and Dribble in their work. Pass and Shoot differ based on whether the ball has reached the goal post or not, but not in terms of actions [24]. Following [24], we instead use six fine grained actions as our classes.

3 Soccer-8k Dataset

With no standard Soccer datasets available, for action recognition in soccer, we started with the creation of a separate dataset. With the help of Event Detector module of our framework (Fig. 1), we extracted one second video clips from input videos and created the Soccer-8k dataset. We took full match Laliga and few Champions League full HD videos available on YouTube for the creation of our soccer action recognition dataset. We extract event information in the video in order to localize the area we present to the classifier. Event is said to happen when a ball is released or gathered in open play. Fig. 2 shows examples of correct event detections. Identification and tracking of both the ball and player in an input video are two sub tasks involved in solving the problem of event detection. Given we have only 2D information, occlusion and the mix up of ball with the players along with blur in video can lead to wrong event identification. Fig. 3 shows examples of incorrect event detections, which are discarded.

It is required of us to first identify the region of video which relates to player's action i.e. who is in contact with the ball and then labeling each of those before passing them as an input to the model. Since it is not easy to manually identify the clips relating a player's action with the ball, we automated the process using ball tracking and player tracking in soccer. However, changes in the camera angles, occlusion of the ball, camera motion and the mix up of ball with the players posed serious challenges.

Handling all those issues, we built the Event Detector module on top of these steps to generate clips that contain only the players' actions with the ball. Eight thousand of such generated clips from special input videos are saved which formed the Soccer-8k dataset.

1. We begin with identifying the relevance and usefulness of each frame. Generally broadcasters broadcast clips that contain closeup view of players or replays of a particular incident from different views for better user experience. Such frames are deemed irrelevant and discarded. Based on predefined norms such as the number of players in the frame, their sizes and play area shown, the frame is analyzed on the visual information contained. Depending on relevance of the frame, ball tracking or ball identification is performed.
2. We define a frame as Pivot frame if the ball is in contact with a player in the current frame, but not in the preceding or succeeding frames. A Pivot Frame

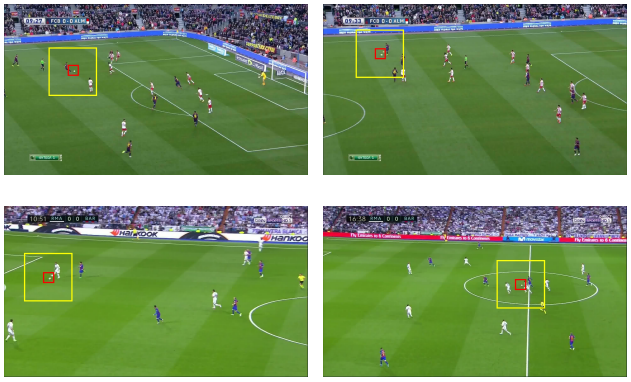

Fig. 2: Examples of correct event detection. The inner red box represents the area where ball has been detected to be in contact with the player. Yellow box represents the area (300×300) around the event detection point that is to be cropped while creating localized video clip from clip generated by stacking frames around the pivot frame.

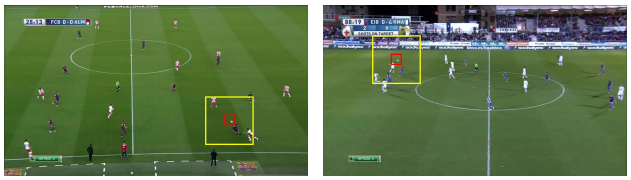

Fig. 3: Examples of incorrect event detection. In few cases, because of incorrect depth perception of the image, the system wrongly detects the occurrence of an event. All such incorrect samples are filtered out during the labeling process.

is labeled as Type-1 if an action takes place in succeeding frames. Else it is labeled as a Type-2 Pivot frame. In general, Type-1 frames occur when the ball has just been received by a player and Type-2 occur when the ball has been released by a player. Fig. 4 shows examples of Pivot Frames.

3. Each relevant frame is checked for Pivot Frame. If P_i is pivot frame, then all the frames in between P_{i-n} and P_{i+m} capture the action that took place in between. It has been observed that most of the actions in soccer games have a span of no more than 24 frames. So we choose n, m as 9, 14 respectively, if the pivot frame is identified as Type-1 and 14, 9, if it is a Type-2 frame.
4. The frames in between P_{i-n} and P_{i+m} are then extracted and a clip is created using them.
5. Off ball player's movements which have no impact on the action can affect the classifier. Hence, we crop $p \times p$ area in video clip centering around the point of event detection forming a new localized video clip that has to be classified. To find the most suitable value for p, we began with low values of p (for better localization of a specific action). Point of event detection is

Fig. 4: Examples of randomly sampled frames from our dataset

shifted in consecutive frames because of factors like camera movement. This results in the failure to capture the whole action. Hence, we had to leverage on localization for camera movement and based on the experiments carried out, value of 300 for p seemed to work well for all the clips generated.

6. The saved clips for dataset are outsourced to soccer enthusiasts who have labeled the actions in each clip. For every sample, the label agreed by at least 50% of the labelers is assigned. Else, the sample is discarded.

The Soccer-8k dataset is created using the above labeled clips. The dataset contains a total of 7942 action clips. We train our model on this dataset, splitting the dataset into training and test sets. To support future work in this area, we made Soccer-8k dataset publicly available.

The table below details the classes and the total number of samples in each class.

Class	**Number of Samples**
Short Pass	2622
Long Pass	670
Header	87
Trapping	1946
Dribbling	1991
Turning	626

Passing implies keeping possession of the ball by maneuvering it on the ground between different players with the objective of advancing it up the playing field. There are 2 types of passes, long and short. A long pass is an attempt to move the ball a long distance down the field via a cross, without the intention to pass it to the feet of the receiving player. A short pass involves keeping the ball low and making it easier for a team mate to control. Header is technique that is used to control the ball using the head to pass, shoot or clear. Trapping refers to intercepting the ball and controlling it. Dribbling is maneuvering of the ball by a single player while moving in a given direction, avoiding defender's attempt to intercept ball. Turning refers to instantaneous shift in player's direction along with the ball to move away from defense [25]. These classes are chosen based on the individual techniques a player should posses in Soccer. [24].

4 Classifier and Implementation Details

GAWAC is a Convolutional Neural Network where in which instead of using a 2D kernel for convolution, 3D kernel is used. This helps in capturing the temporal dependencies between the frames of a video along with spatial dependencies.

GAWAC (Fig. 5) has 5 convolution, 4 max-pooling, and 2 fully connected layers, followed by a softmax output layer. After the fourth convolution layer, we use a Gaussian distributed weight filter to emphasize the information at the center. We do this to make the model learn that the action is concentrated around the ball, centering which the clip is extracted. This helps us in removing irrelevant information about other players in the 300×300 image.

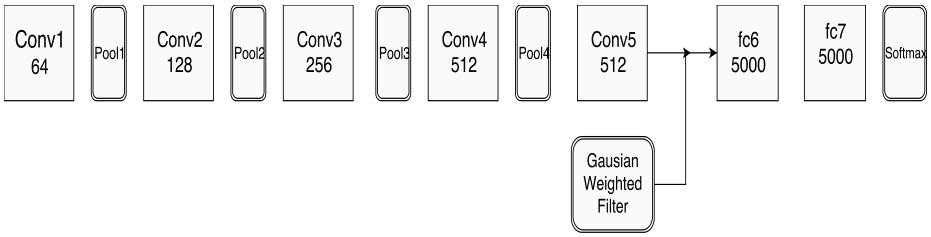

Fig. 5: The GAWAC architecture

All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters in convolution layers are 64, 128, 256, 256 and 512 respectively. Except for the first maxpool layer having dimensions $1 \times 2 \times 2$, all other maxpool layers have dimensions $2 \times 2 \times 2$. Each fully connected layer has 5000 output units. The input dimensions are $3 \times 12 \times 100 \times 100$. Fig. 1 presents the work flow of the model.

5 Experiments on Soccer-8k dataset

The experiments are conducted on our Soccer-8k dataset. It contains 7942 videos of six different soccer actions performed by different players in different matches at different venues. The six actions are:

- Short Pass
- Long Pass
- Heading
- Trapping
- Turning
- Dribbling

All of these videos are RGB and have moving backgrounds which is one of the challenges. All these video clips have 24 frames. Fig. 11 - 16 shows sample frames from few video clips.

For each experiment, 70% of Soccer-8k dataset is reserved as training set and rest 30% is divided equally as test and validation datasets.

Imbalance in number of samples of each class may lead to biased classification. To handle this we up-sample the classes with low number of samples ensuring equal number in all classes. To avoid repetition the same example to network we augment the data which ensures better learning, reducing the chances of over-fitting.

C3D model is the best performer on UCF101 dataset for video-based sport identification, which is rich in visual content including video collections of Baseball Pitch, Basketball Shooting, Diving and etc. [26] has proved that LSTMs are very suitable to learn the long-term time dependencies in between the individual frames of a video. Hence we use the state-of-art C3D and LSTM architectures for comparison with our GAWAC model. All models are implemented in keras [27] deep learning framework.

5.1 Experiments on GAWAC

The experiments on GAWAC model were run for different Gaussian distributions. We tried different Gaussian distributions, with Mean (μ) and Variance (σ^2) as (0, 1.0), (0, 0.5) and (0, 0.2). These experiments were run on Nvidia GeForce GTX 1080 Ti GPU, with 11GB of VRAM. With batch size as 32, the experiments are run on the augmented dataset with approximate size of 15,000 samples. The model is trained on new augmented dataset which gets generated on the fly in each epoch. This ensures that the model never gets the same example in training. Techniques such as shifting, rotation, centering, flip etc., are used to generate augmented video samples. The input clip is resized from $12 \times 3 \times 300 \times 300$ to $12 \times 3 \times 100 \times 100$ and inputted to the model. Here 12, 3, 100 and 100 denote number of frames in the clip, number of channels in each frame, height and width of video respectively.

To avoid over-fitting, early stopping was used as criterion, which is determined from validation accuracy. Stochastic Gradient Descent (SGD) optimizer

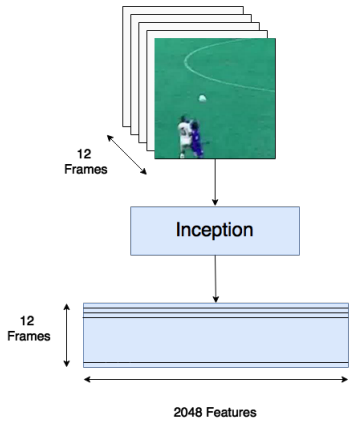

Fig. 6: Block diagram of the Inception model

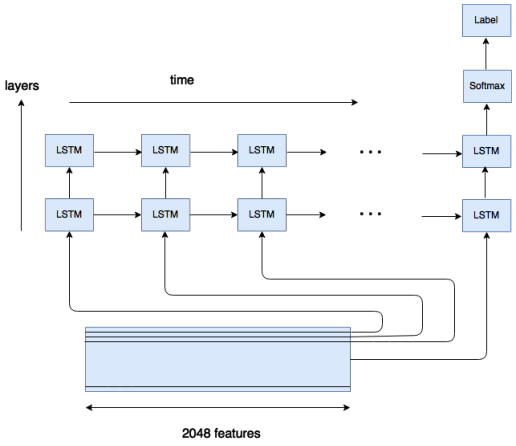

Fig. 7: The 2 layered LSTM architecture, which works on the features obtained from the Inception model

with Nestrov Momentum parameter set to 0.9 was used to train the model. Initial Learning rate was set to 0.005, with a decay of 1e-4. All other parameters are set to the default values of SGD optimizer in keras. To avoid over-fitting, dropout of 0.8 was introduced between fully connected layers of network, L2 weight regularization was used with lambda of 0.0001 in all the convolution layers and 0.00005 in fully connected layers of network.

5.2 Experiments on C3D

All the experiments on C3D were run on same stack with similar settings and hyper parameters.

5.3 Experiments on LSTM architecture

For this model, with data augmentation, a total of 100,000 samples are generated with equal number of samples in each class. We use shifting, rotation, centering, flip etc as techniques for augmentation. This dataset is fed as input to Inception model (Fig. 6) and features were extracted from the penultimate layer. These features are fed to two layered LSTM (Fig. 7) that has 400 LSTM cells in each layer.

All the experiments were run on the same stack. With the same batch size i.e. 32, the experiments are run on the new augmented dataset of 100,000 samples. To avoid over-fitting, early stopping was used as criterion, which is determined from validation accuracy. SGD optimizer was used for training this model. The momentum value here is also set as 0.9. Initial Learning rate was set to 0.01, with decay of 1e-4. All other parameters are set to default SGD values in keras.

In order to avoid over-fitting, we introduced dropout of 0.5 between all the layers and L2 regularization with lambda of 0.0001 was used in all the layers of the network.

6 Results

Figures 8, 9 and 10 presents the confusion matrices for GAWAC, C3D and 2D LSTM models respectively, for Soccer-8k dataset. Short pass and long pass differ slightly with respect to the action of the player. The differentiation between the two actions becomes difficult if the motion of the ball is not noticed. This results in confusion in the models. Though the models are expected to perform well in Header class, the very few samples in the dataset resulted in improper learning for this class. Similar to Long and Short passes, the act of Turning is much similar to that of Trapping and Dribbling, and hence higher confusion in case of Turning samples.

Moreover, the common actions that are possible in every sample, such as running, jumping can impact the decision to an extent. This can often result in wrong classification of the sample. Cropping the video to a size of 300×300 around the point of event with an aim to capture the entire action has however resulted in the capturing of unnecessary information in the samples. This can have significant effect on the features learnt. Addition of Gaussian filter emphasized on the learning features around the ball and hence accuracy is higher in GAWAC.

The table below conveys the F1 scores of each model.

Model	**F1 Score**
C3D	46.02
Inception + LSTM	41.88
GAWAC with $\sigma^2 = 1.0$	55.72
GAWAC with $\sigma^2 = 0.5$	62.75
GAWAC with $\sigma^2 = 0.2$	52.8

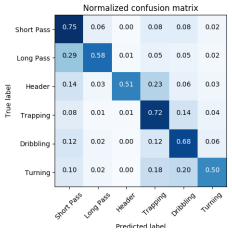

Fig. 8: Confusion Matrix of GAWAC trained on Soccer-8k, $\sigma^2 = 0.5$

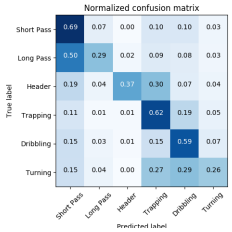

Fig. 9: Confusion Matrix of C3D trained on Soccer-8k

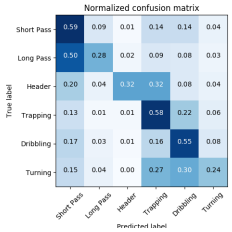

Fig. 10: Confusion Matrix of 2 layered LSTM trained on Soccer-8k

7 Conclusion and Future Work

We have presented a novel event based framework for fine grained action recognition in the sport of Soccer, *which is first of its kind*. We compared our model with the state-of-art techniques C3D and LSTM. The models are trained on the newly created custom dataset, Soccer-8k, that contains approximately 8000 samples and a total of 6 classes. In addition, our framework can recognize actions of players from live feed and can be used for stats generation. With slight modification, our framework can be generalized to other multi-player sports too. Moreover, the errors made by the models are because of the similarity in classes implying that the features they have learned are significant and are semantically meaningful.

This work can be extended for much finer action recognition, such as classifying an action into subtypes. For example a pass can be classified as either left or right footed pass or either outside or inside foot pass. This work opens new dimensions in analysis of soccer games without any need of additional infrastructure like electronic chips and can be afforded even by low budget clubs. This work with addition of multi player tracking and identification can be used for generation of textual commentary on fly and for text query based video search. Player statistics, team statistics and coordination among the team members are few things that can be extracted and analyzed.

References

1. Chen, L., Hoey, J., Nugent, C.D., Cook, D.J., Yu, Z.: Sensor-based activity recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42** (2012) 790–808
2. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35** (2013) 221–231
3. Tsunoda, T., Komori, Y., Matsugu, M., Harada, T.: Football action recognition using hierarchical lstm. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE (2017) 155–163
4. Mora, S.V., Knottenbelt, W.J.: Deep learning for domain-specific action recognition in tennis. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE (2017) 170–178
5. Parmar, P., Morris, B.T.: Learning to score olympic events. arXiv preprint arXiv:1611.05125 (2016)
6. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. (2014) 568–576
7. Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M.: A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1961–1970
8. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. arXiv preprint arXiv:1611.09078 (2016)

Fig. 11: An example of short Pass sample

Fig. 12: An example of long Pass sample

Fig. 13: An example of Turning sample

Fig. 14: An example of Trapping sample

Fig. 15: An example of Dribbling sample

Fig. 16: An example of Header sample

9. Chen, S., Feng, Z., Lu, Q., Mahasseni, B., Fiez, T., Fern, A., Todorovic, S.: Play type recognition in real-world football video. In: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, IEEE (2014) 652–659
10. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1971–1980
11. Maksai, A., Wang, X., Fua, P.: What players do with the ball: A physically constrained interaction modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 972–981
12. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision. (2013) 3551–3558
13. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 4694–4702
14. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1933–1941
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
16. Zhang, B., Wang, L., Wang, Z., Qiao, Y., Wang, H.: Real-time action recognition with enhanced motion vector cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2718–2726
17. Feichtenhofer, C., Pinz, A., Wildes, R.: Spatiotemporal residual networks for video action recognition. In: Advances in Neural Information Processing Systems. (2016) 3468–3476
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1., IEEE (2005) 886–893
19. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
20. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision **103** (2013) 60–79
21. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. Computer Vision–ECCV 2010 (2010) 143–156
22. Gourgari, S., Goudelis, G., Karpouzis, K., Kollias, S.: Thetis: Three dimensional tennis shots a human action dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2013) 676–681
23. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. (2017) 4278–4284
24. AS, S.: 50 selected soccer skills and drills (2003)
25. Wikipedia: Association football — wikipedia, the free encyclopedia (2017) [Online; accessed 1-December-2017].
26. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9** (1997) 1735–1780
27. Chollet, F., et al.: Keras (2015)