Football match intelligent editing system based on deep learning

Bin Wang[1], Wei Shen[1,2], FanSheng Chen[3] and Dan Zeng[1*]
[1] Key Laboratory of Specialty Fiber Optics and Optical Access Networks,
Joint International ResearchLaboratory of Specialty Fiber Optics and Advanced Communication,
Shanghai Institute of Advanced Communication and Data Science, Shanghai University,
Shanghai, 200444 - China
[e-mail: wangbin418@outlook.com, wei.shen@t.shu.edu.cn, dzeng1993@shu.edu.cn]
[2] Department of Computer Science, Johns Hopkins University,
Baltimore, MD 21218 - USA
[e-mail: shenwei1231@gmail.com]
[3] Key Laboratory of Intelligent Infrared Perception , Chinese Academy of Sciences,
Shanghai, 200083 - China
[e-mail: cfs@mail.sitp.ac.cn]
*Corresponding author: Dan Zeng

Received December 26, 2018; revised February 13, 2019; revised March 19, 2019; accepted April 6, 2019; published October 31, 2019

Abstract

Football (soccer) is one of the most popular sports in the world. A huge number of people watch live football matches by TV or Internet. A football match takes 90 minutes, but viewers may only want to watch a few highlights to save their time. As far as we know, there is no such a product that can be put into use to achieve intelligent highlight extraction from live football matches. In this paper, we propose an intelligent editing system for live football matches. Our system can automatically extract a series of highlights, such as goal, shoot, corner kick, red yellow card and the appearance of star players, from the live stream of a football match. Our system has been integrated into live streaming platforms during the 2018 FIFA World Cup and performed fairly well.

Keywords: live football match, intelligent editing system, highlight extraction, deep learning, object detection

http://doi.org/10.3837/tiis.2019.10.017 **ISSN : 1976-7277**

1. Introduction

Video analytics technology has been extensively studied to provide users with faster and more convenient access to interesting or important parts of the video. In particular, with the explosion of multimedia videos, the demand for high-performance image and video indexing and retrieval technology has become greater and greater [9,10,11,12]. Among them, the video summary is considered to be very critical. Users only need to watch some important parts to save time. Manually analyzing and summarizing video sequences is a laborious and labor-intensive task. Due to the large number of available sequences and the amount of time required in the process, it is desirable to provide an automatic motion video sequence highlighting method in particular.

One of the most widely studied video types is sports game because it has a large audience and has more regular features than some other videos. Specifically, for sports video, an effective abstraction approach is the highlighting technique, which focuses on how to generate a summary of the match that includes all of its interesting parts. Some previous researchers have proposed a number of highlighting methods for general sports competitions and specific types of sports matches. Ekin et al. [13] detected play and break events in sports videos to generate a summary. Some other researchers used the slow-motion replay to summarize sports videos [14,15]. However, due to the differences and diversity of different games, analyzing general sports games remains an open issue. Some researchers turned to study specific sports, such as football, basketball or diving. This paper focuses on the techniques for football video highlights extraction. Ancona et al. [16] proposed a target detection method using SVM classifiers in soccer videos. Zawbaa et al. [17] proposed a system that first splits the entire video sequence into small video shots and then classifies the resulting shots into different shot types. After that, the system applies SVM and an artificial neural network algorithm to select segments with special performance. Subsequently, the system then detects the vertical goalpost and goal net. Finally, the important events during the game will be highlighted in the football video summary. Fendri et al. [18] proposed a segmentation and index based approach that relies on low-level and text-based processing of soccer videos. Ekin et al. [19] proposed a fully automated and computationally efficient framework for analyzing and summarizing football videos using cinematic and object based features. Lofti et al. [20] proposed a way to reject the shots containing non-significant events to summarize the video. Tabii et al. [21] introduced a new method based on lens detection, lens classification and the finite state machine technology to automatically extract soccer video summaries. All of the above methods use artificially designed features to process video which lacking the generalization ability, and are difficult to put into practical use. R Yan et al. [33] explore a new "One to Key" idea to progressively aggregate temporal dynamics of key actors with different participation degrees over time from each person to improve group activity recognition performance. It's smart and efficient, but it doesn't distinguish the highlights of the game. With the development of artificial intelligence technology, people's lifestyle is also constantly automated and intelligent, and the future multimedia processing technology will be intelligent and efficient.

In this paper, we present an intelligent editing system based on deep learning to generate highlights of football game videos or live streams and solves the problems deep learning brings. We do not need to manually design feature extraction techniques like previous methods, and we do not need to replace algorithms for different scenarios. The system can generate a variety of interesting short videos in real time, such as shooting, penalty kicks, goals,

corner kicks, red and yellow cards, etc., and quickly generate star clips after the end of the live broadcast. The system achieves 100% recall in the 2018 FIFA World Cup.

2. Overview of the system

We develop an intelligent editing system to automatically extract highlights from live football matches. In this system, first, the live stream of a football match is passed through different modules, including shot segmentation, red and yellow card detection, corner kick detection, penalty kick detection, shoot and celebration detection, score detection and face detection, to produce shot boundary frames, special action frames and star player frames. Then, the information is fused to produce highlights in an integration module. **Fig. 1** shows the flow chart of our system.

To build such a system, we prepare a large number of images from football match videos such as 2014 Brazil World Cup and annotate the objects we need, such as players, red cards, yellow cards, corner flags and footballs.

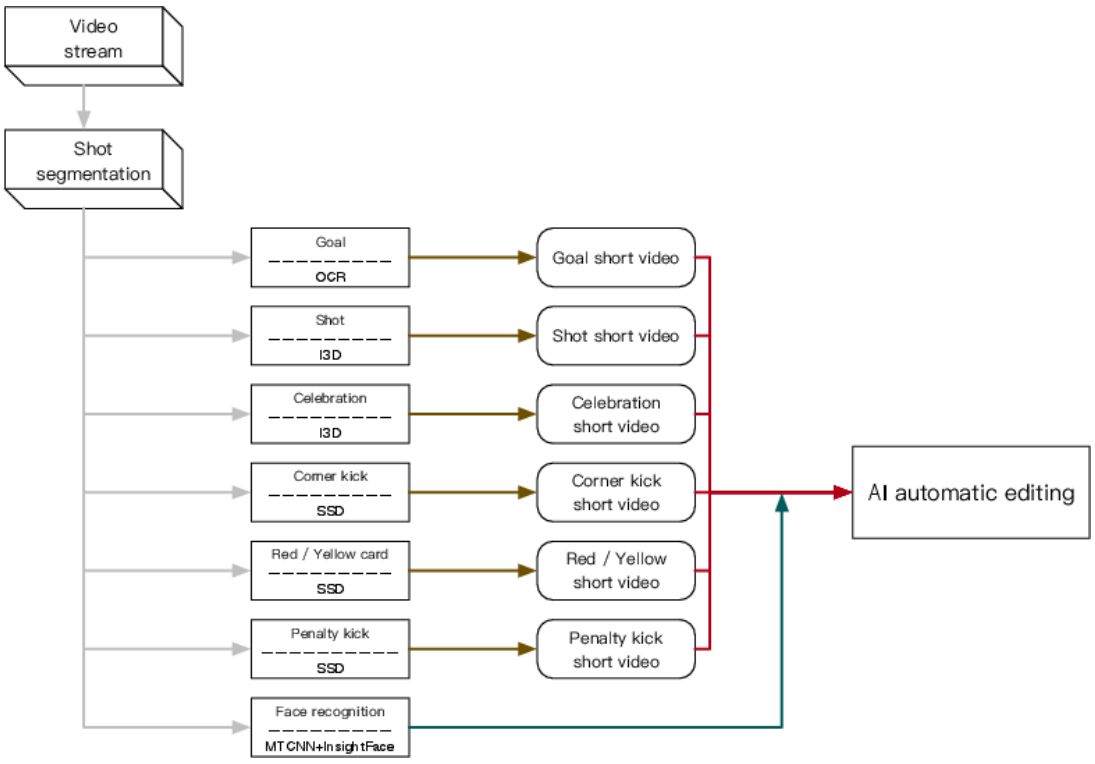

Fig. 1. The overall system. The live stream is fed into all modules.

3. Method

This section specifically introduces the details of each module in our system.

3.1 Shot Segmentation

Shot segmentation is the basic module of our system because highlights are generated by

synthesizing continuous shots to prevent discontinuity. Our goal is to find shot boundary frames, such as the one shown in **Fig. 2**.

After comparing different techniques for shot boundary detection [1], we use a histogram method to segment videos, which can avoid the differences caused by the movements of objects in the scene, especially in football match videos. The simplest histogram method compares gray or color histograms of two successive frames. A shot boundary is then detected by thresholding the bin-wise difference between the two histograms. Ueda et al. [2] used the color histogram change rate to find shot boundaries and achieved good results. However, they just use the absolute value ratio of the color histogram between two frames, so the result is not sensitive. There are a lot of missed detections in **Table 2**.

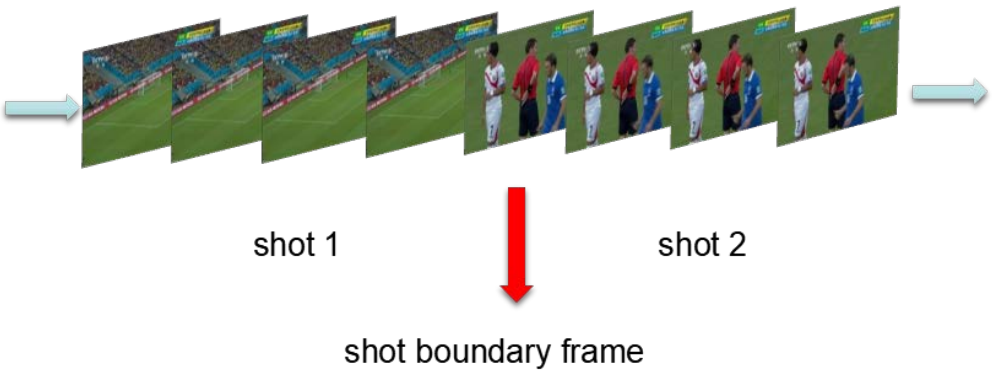

Fig. 2. Shot segmentation.The left and right sides of the red arrow are different shots.We define the first frame of each shot as a boundary frame.

We assume that the difference D_{cur} between a shot boundary frame and its previous frame should be greater than the average D_{avg} of all the frames in the shot. D_{cur} and D_{avg} are computed by:

$$D_{cur} = \sum \frac{(h(c)-h(c-1))^2}{\max(h(c),h(c-1))+\epsilon}, c \geq 2 \quad (1)$$

$$D_{avg} = \frac{\sum_{i=2}^{c-1} D_{cur}(i)}{c-2}, c \geq 3 \quad (2)$$

where h(·) is a function to calculate the histogram and c is the frame index. ϵ is an arbitrarily small positive quantity to prevent the denominator being 0. We set it to 0.0001. We take the cth frame as a shot boundary frame when the ratio of D_{cur} to D_{avg} is above a threshold. The threshold is obtained by cross-validation on a set of football match videos.

3.2 Red/Yellow Card Detection

A red or yellow card is a type of penalty card that is shown in many sports after a foul. The yellow card is a serious warning and the red card indicates an exit. In a match, players who accumulate two yellow cards will immediately receive a red card.

Fig. 3. Red/yellow card detection. This module detects the red and yellow cards in a frame, as the red and yellow boxes in the figure.

To detect red and yellow cards in football match videos, we need to use object detection algorithms. Object detection algorithms are mainly divided into three branches, namely multi-stage, two-stage and one-stage. Early methods, such as R-CNN [22] and SPPNet [23], are the multi-stage methods. Selective Serach, Feature extraction, location regression, and classifier are divided into multiple stages to be individually trained. When it comes to Fast R-CNN [24], Feature extraction, location regression, and classifier are all integrated into a network, and these three tasks can be trained together. Since the task that generates the region proposal needs additional training, it is called two-stage. Later, the one-stage network of the Yolo series [25, 26, 27] and SSD [7] greatly improved the speed of object detection by removing region proposal network.

In order to balance speed and accuracy, we choose SSD as the framework of our object detection algorithm. We perform SSD on each frame of the football match video, and then output those frames where red or yellow cards are detected, as shown in **Fig. 3**.

3.3 Corner Kick Detection

A corner kick is the method of restarting the play in a football game when the ball goes out of the goal line, without a goal being scored, and having been last touched by a member of the defending team. Since corner kicks are considered to be a good goal scoring opportunity for the attacking side, many viewers enjoy watching highlights of this aspect.

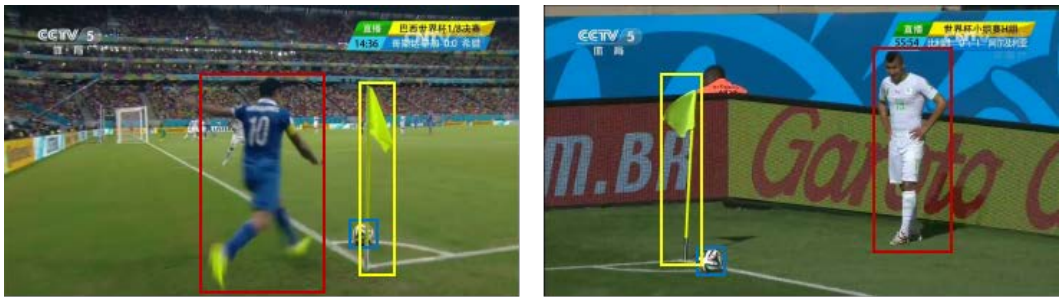

Fig. 4. Corner kick. This module detects the players, the ball and the corner flag in a frame, as the red, blue and yellow boxes in the figure.

Corner kick detection is also performed by SSD. Considering that in a corner kick frame, the player, the ball and the corner flag should appear at the same time, as shown in **Fig. 4**. We use the trained SSD to process frames to find what is in the frame. When one player, the ball and the corner flag are detected simultaneously in one frame, we output this frame as a

keyframe of corner kick.

3.4 Penalty kick detection

A penalty kick is a method of restarting a play in the football game, in which a player is allowed to take a single shot on the goal while it is defended only by the opposing team's goalkeeper. It is awarded when a foul punishable by a direct free kick is committed by a player in his or her own penalty area.

SSD is also used in this module. We need to detect the position of the player and the goal, as shown in **Fig. 5**. In a frame, when the number of players on one side of frame is over ten and the goal appears on the other side, this frame is considered to be the frame in the penalty kick shot and output.

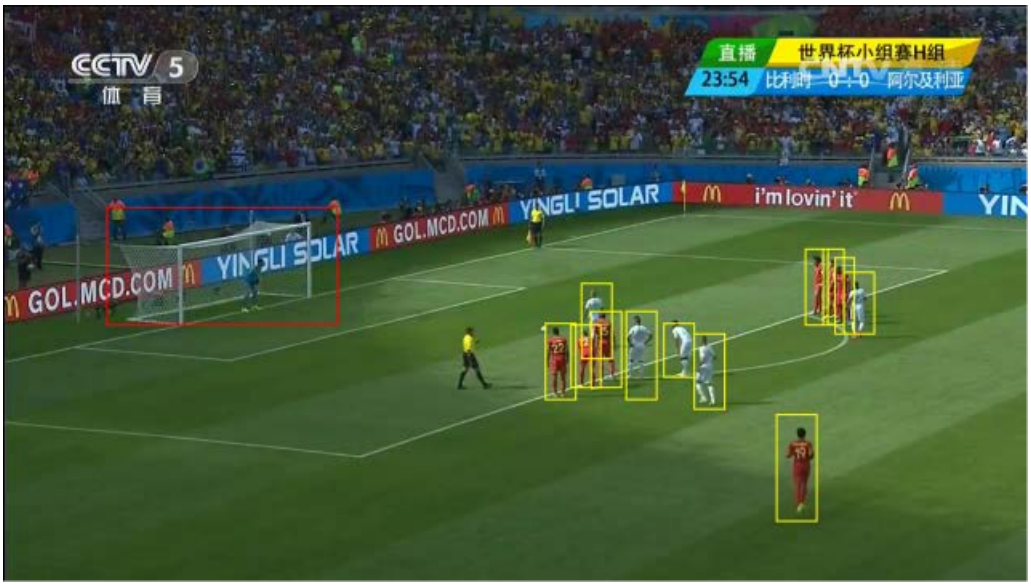

Fig. 5. Corner kick. This module detects the players and the goal in a frame, as the yellow and red boxes in the figure.

3.5 Shoot and Celebration Detection

The shoot is the climax of the football game. Since every shot will provide a chance to score, it attracts the attention of the audience. The celebration action also marks the occurrence of important events.

In this module, we need to detect the shoot and celebration from the live football matches. Shooting and celebration are momentary actions but we require continuous frames to find them in the video. We formulate shoot and celebration detection as an action recognition problem, whose goal is to determine the type of human behavior in a video.

IDT [28] is the most stable and reliable method before deep learning has been widely applied to the field of action recognition, but this algorithm is very slow. It uses the optical flow between the two frames of video and the SURF key points to match, thereby reducing the effects of camera motion. With the development of deep learning, many action recognition methods based on deep learning have been proposed. The Two Stream Network [29,30] computes a dense optical stream for every two frames in the video sequence, and then trains

the CNN model for the video image and the dense optical stream respectively. The two branches of the network respectively judge the action categories, and finally directly fuses class scores of two networks to get the final classification result. C3D [31] uses 3D convolution and 3D Pooling to build a network to handle video directly. I3D uses a better architecture and is pre-trained on the more generic Kinetic dataset. P3D [32] uses experiments to prove that time domain information is important for action recognition.

We adopt the well-known action recognition model I3D [8] for shoot and celebration detection and experiment results are shown in **Table 3**. The number of shoot and celebration in a game is limited, which requires us to do data augmentation. We take a clip containing shooting or celebration and its adjacent clips as positive samples. Sufficient training data is generated by randomly offsetting these positive samples back and forth in the time dimension and I3D will be trained on them. In the testing phase, a live stream is fed into the module and the module outputs the middle frame's position in this live stream that was detected as shooting or celebration, as shown in **Fig. 6**.

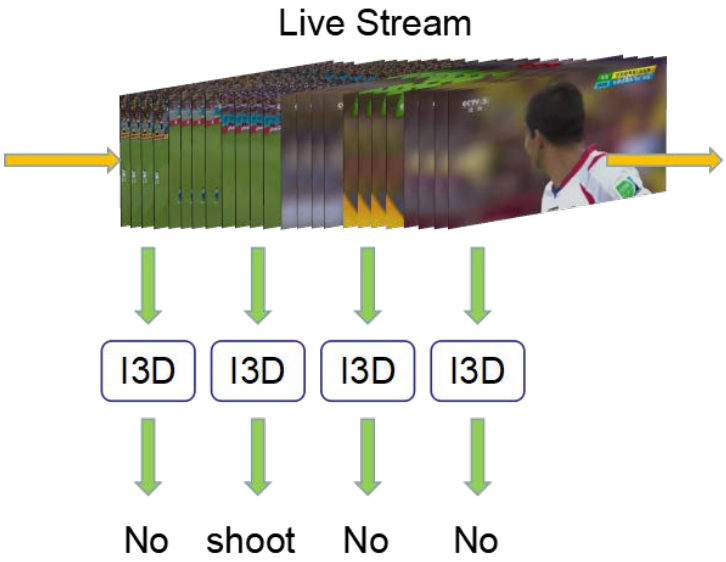

Fig. 6. Shoot and celebration detection. The live stream is split into a number of different short videos, which are then inputed into I3D to predict whether there is a shoot or celebration action.

3.6 Score Detection

Goal is the most exciting part of a game and is one of the important components of highlights. We design a score detection system, which detects goals by detecting the change of the score. This system contains three components: text region detection, text recognition and post-processing.

For text region detection, we use CTPN [3]. We train this model on a large number of text images, to detect text regions in each frame of football match videos. However, in the frames of football match videos, the score takes up a very small area, leading to unstable results. Therefore, we propose a simple strategy to make better use of CTPN to address the size imbalance between the score area and the input frame. We first find the text regions containing scores, and then input the obtained regions into CTPN again to get more precise results.

Formally, we input a frame c into CTPN to get text regions which we call them boxes $\{B_1(\mathrm{c}), B_2(\mathrm{c}), \ldots, B_N(\mathrm{c})\}$, where N_c is the number of boxes detected in this frame. Since a score area should remain unchanged in adjacent frames, we take these regions in adjacent frames as candidate score regions, if the Intersection over Union (IoU) between them is greater than a threshold. The IoU between two regions in different frames is defined as follow:

$$Iou_{i,j}(c) = \frac{B_i(c) \cap B_j(c-1)}{B_i(c) \cup B_j(c-1)}, c > 1 \tag{3}$$

$Iou_{i,j}(c)$ represents IoU between the ith box of the cth frame and the jth box of $(\mathrm{c}-1)$th frame. Then we will extend the range of these candidate regions and extract them from frame c. In our system, we have expanded 150 pixels and 100 pixels in length and width, respectively. Finally, we input these candidate regions into CTPN again to get a set of score regions.

For text recognition, we use OCR technology to recognize words in text regions. We adopt Tesseract [4] as our method because it is mature and free. The output of this component is a sequence of all the detected words.

Now, we have a sequence of words which contains scores. Next, we need to extract the scores from this sequence by post-processing. Regular expressions are perfect for this task. It can find words matching the format of scores in the sequence. We record scores and confirm a goal frame when there is a legitimate change in the score like zero colon zero to one colon zero.

3.7 Face Recognition

In this module, we want to detect soccer stars from soccer match videos by face recognition. However, this is a challenging task, because in soccer match videos, faces are blurred and most of them are profiles. Most current face recognition training datasets are front faces. Such data bias inevitably results in a performance drop.

The face recognition model we use is ArcFace [5], without fine-tuning, because the performance of the existing model is good enough. The face detection model we used is MTCNN [6], a very fast model with promising performance.

Image registration should be prepared before face recognition. Registration images are usually recent front face photos of the person to be recognized, but we cannot collect the recent front face images of soccer stars. We believe that using images in soccer match videos instead of using web images is more appropriate to the application scenario of soccer matches. However, it is time-consuming and laborious to find star faces directly in soccer match videos, thus we propose a two-step approach. The first step is to use web image as registration images to find soccer stars in the soccer match videos in recent two years. We search for front face images of these stars on the Internet and pre-process them as registration images. For match videos, we perform face detection on one frame per five frames, then perform the same pre-processing method used for registration images on the detected face regions. After pre-processing, we feed the detected face regions into ArcFace to do face recognition. Since training images and the testing images in match videos are quite different, we set a lower threshold to ensure that as many star face images as possible are captured. In the second step, we manually search the clear front face image for each star obtained in the first step. The number of registration images of each star is approximately 15.

The test phase uses the previously prepared registration images. The output of the face recognition module is which frame contains which star and where the star face is.

3.7 Integration module

The above described modules output the positions of key frames of highlights and the positions of the shot boundary frames. We combine a shot containing a key frame as well as its nearby shots to create a short video which lasts about a minute. These combined short videos are the highlights output of our system. Just like **Fig. 1**.

4. System Details

In the testing phase, in order to ensure that the entire system is real-time, each module need its own settings to increase the speed.

Shot Segmentation module is used on every frame. Red and yellow card detection and corner kick detection are performed every eight frames from the input stream. In the shoot and celebration detection, we take one frame out of every eight frames from the live stream and combine them into a short video of 2.56 seconds long. These short videos will be input into I3D. Score detection, face detection and recognition are performed every 25 frames and every 10 frames, respectively.

5. Experiments

5.1 Shot Segmentation

Three 2014 Brazil World Cup match videos are used to verify the shot segmentation module of our system: Costa Rica vs Greece, Portugal vs Ghana and Belgium vs Algeria. All videos are 720p and the first 10,000 frames of each video are used. The statistics of the test data are shown in **Table 1**. We compare three histogram methods in **Table 2**, where Abs [2] calculated the absolute value instead of the square. “Correct” is the number of correctly detected shot boundary frames, “False Positive” is the number of wrongly detected frames and “Missed” indicates the number of shot boundary frames which are fail to detect. As can be seen from the **Table 2**, our method is better than [2] in terms of both accuracy and stability.

The selected application lists for each class and the number of applications in each class are shown in **Table 1**.

Table 1. The statistics of the data used in experiments.

Video Name	number of frames	cuts
Costa Rica vs Greece	10000	40
Portugal vs Ghana	10000	31
Belgium vs Algeria	10000	55
Total	30000	126

Table 2. Shot segmentation results.

Method	Correct	FalsePositive	Missed
Abs[2]	83	8	43
W/O Avg(ours)	94	111	32
Avg(ours)	116	28	10

5.2 Shoot and Celebration Detection

We compared the C3D, P3D and I3D in a football match video. We extract 1000 clips from a video, each of which is 2.56 s long and contains 16 frames. In other words, we take a frame from every 4 frames. Among the 1000 clips, there are 13 shoots, and 5 celebrations. The results are shown in **Table 3**. We will filter out the results with a confidence level below 0.5. From the experimental results, I3D can find all the required actions in the test samples, but the number of misjudgments is slightly higher than C3D. We pay more attention to whether the model can detect all the special actions, so we choose I3D as the model for shoot and celebration detection.

Table 3. Shoot and Celebration Detection results.

Method	Correct	FalsePositive	Missed
P3D[33]	2	101	15
C3D[32]	11	14	6
I3D[8]	17	33	0

5.3 Score Detection

We tested CTPN and refined CTPN on a short match video and compared their results. The results are shown in **Table 4**. The accuracy is measured by the ratio of the number of correctly identifying scores to the number of occurrences of the score. The accuracy of our refined method is much higher than the original method without refined.

Table 4. Score detection results.

Method	Accuracy
W/O Refined	0.062
Refined(ours)	0.882

5.4 Face Recognition

In order to verify the performance of the face recognition module, we identify eight stars in the video of two Brazilian World Cup matches. The results are shown in **Table 5**. Our system can achieve more than 98% recognition accuracy for most stars. These stars can be correctly identified even in very complicated scenes, as shown in **Fig. 7**.

Table 5. Experiment of Face Recognition.

Game Name	Player Name	Correct	FalsePositive	Accuracy
Spain Vs Chile	Ramos	524	10	98.13%
	Pique	311	5	98.42%
	de Gea	91	8	91.92%
	Silva	114	0	100.00%
Croatia Vs Mexico	Hernandez	544	11	98.02%
	Manjukic	308	34	90.06%
	Rakitic	91	9	95.26%
	Modric	114	1	99.31%

Fig. 7. Face detection and recognition in the 2018 FIFA World Cup. Some examples of correct detection and recognition in complex scenes such as side faces, occlusion and blurring are shown here.

5.5 Entire System

Our system was put into use during the 2018 World Cup. We define short videos that include shoot, red and yellow card, corner kick, penalty kick, goal, and celebration as highlights. Our system achieved 100% recall during the game but with a not high accuracy. Manual filitering is required before uploading short videos to the network, so recall is more important than precision.

5. Conclusion

In this paper, we developed a novel system for automated highlight extraction in soccer videos or football match lives. This system consists of deep learning based action recognition, objects detection and face recognition modules. We first utilized color histogram features to extract shot boundary frames. Then deep learning techniques are used to find the keyframes of the highlights and star shots. Finally, the wonderful short videos and star clips are generated by splicing shots containing keyframes. The experimental results show the promising performance of the various modules of this system. Our system also has been integrated into live streaming platforms during the 2018 FIFA World Cup and performed very well.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61572307).

References

[1] Boreczky J S, Rowe L A, “Comparison of Video Shot Boundary Detection Techniques,” *Journal of Electronic Imaging*, 5(2), 32-38, 1996. Article (CrossRef Link).
[2] H. Ueda, T. Miyatake, and S. Yoshizawa, “IMPACT: an interactive natural-motion-picture dedicated multimedia authoring system,” in *Proc. of CHI, ACM, New York*, pp. 343-350, 1991. Article (CrossRef Link).
[3] Tian, Z., Huang, W., He, T., He, P., & Qiao, Y., “Detecting text in natural image with connectionist text proposal network,” in *Proc. of uropean Conference on Computer Vision*, 56-72, 2016. Article (CrossRef Link).
[4] Smith R, “An Overview of the Tesseract OCR Engine,” in *Proc. of International Conference on Document Analysis and Recognition. IEEE Computer Society*, 629-633, 2007.
[5] Deng J, Guo J, Zafeiriou S, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” 2018.
[6] Zhang K, Zhang Z, Li Z, et al., “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” *IEEE Signal Processing Letters*, 23(10), 1499-1503, 2016. Article (CrossRef Link).
[7] Liu W, Anguelov D, Erhan D, et al., “SSD: Single Shot MultiBox Detector,” in *Proc. of European Conference on Computer Vision. Springer International Publishing*, 21-37, 2016.
[8] Carreira J, Zisserman A, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” *Computer Vision and Pattern Recognition. IEEE*, 4724-4733, 2017.
[9] Hu W, Xie N, Li L, et al., “A Survey on Visual Content-Based Video Indexing and Retrieval,” *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, 41(6), 797-819, 2011. Article (CrossRef Link).
[10] Datta R , Joshi D , Li J, et al., “Image retrieval: Ideas, influences, and trends of the new age,” *Acm Computing Surveys*, 40(2), 1-60, 2008. Article (CrossRef Link).

[11] Lew M S, "Content-based Multimedia Information Retrieval : State of the art and challenges," *Acm Transactions on Multimedia Computing Communications & Applications*, 2(1), 1-19, 2006. Article (CrossRef Link).
[12] Schoeffmann K, Hopfgartner F, Marques O, et al., "Video browsing interfaces and applications: a review," *Spie Reviews*, 1(1), 018004, 2010.
[13] A. Ekin and M. Tekalp, "Generic play-break event detection for summarization and hierarchical sports video analysis," in *Proc. of 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)(ICME), Baltimore, MD, USA*, pp. 169-172, 2003. Article (CrossRef Link).
[14] Pan H, Van Beek P, Sezan M I, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. of Acoustics, Speech, & Signal Processing, on IEEE International Conference. IEEE Computer Society*, 2001.
[15] Pan H, Li B, Sezan M I, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," in *Proc. of IEEE International Conference on Acoustics. IEEE*, 2002.
[16] Ancona, N., Cicirelli, G., Branca, A., Distante, A, "Goal detection in football by using support vector machines for classification," in *Proc. of International Joint Conference on Neural Networks*, Vol. 1, 15-19, 2001. Article (CrossRef Link).
[17] Zawbaa H M, El-Bendary N, Ella Hassanien A, Kim T, "Event detection based approach for soccer video summarization using machine learning," *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 7, no. 2, pp 63-80, April 2012.
[18] Fendri E, Ben-Abdallah H, Ben Hamadou A, "A novel approach for soccer video summarization," in *Proc. of Second International Conference on Multimedia and Information Technology (MMIT 2010), Kaifeng, China*, pp. 138-141, April 24 - 25, 2010. Article (CrossRef Link).
[19] Ekin A , Tekalp A M , Mehrotra R, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, 12(7), 796-807, 2003. Article (CrossRef Link).
[20] Lotfi E, Pourreza H R, "Event detection and automatic summarization in soccer video," in *Proc. of 4th Iranian Conference on Machine Vision and Image Processing (MVIP07), 2007 Mashhad, Iran*, 2007.
[21] Tabii Y, Oulad Haj Thami R, "A new method for soccer video summarizing based on shot detection,classification and finite state machine," in *Proc. of 5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications (SETIT 2009), Hammamet, Tunisia*, pp. 7-11, March 22-26, 2009
[22] Girshick R, Donahue J, Darrell T, et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proc. of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society*, 2014.
[23] He K , Zhang X , Ren S , et al., "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(9), 1904-1916, 2015. Article (CrossRef Link).
[24] Ren S, He K, Girshick R, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6), 1137-1149, 2017. Article (CrossRef Link).
[25] Redmon J, Divvala S, Girshick R, et al., "You Only Look Once: Unified, Real-Time Object Detection," 2015.
[26] Redmon J, Farhadi A, "[IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Honolulu, HI (2017.7.21-2017.7.26)] 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - YOLO9000: Better, Faster, Stronger," 6517-6525, 2017.
[27] Redmon J , Farhadi A, "YOLOv3: An Incremental Improvement," 2018.
[28] Wang H, Schmid C, "Action recognition with improved trajectories," in *Proc. of the IEEE international conference on computer vision*, 3551-3558, 2013.
[29] Simonyan K, Zisserman A, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, 568-576, 2014.

[30] Feichtenhofer C, Pinz A, Zisserman A, "Convolutional two-stream network fusion for video action recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 1933-1941, 2016.
[31] Tran D, Bourdev L, Fergus R, et al., "Learning spatiotemporal features with 3d convolutional networks," in *Proc. of the IEEE international conference on computer vision*, 4489-4497, 2015.
[32] Qiu Z, Yao T, Mei T, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV), IEEE*, 5534-5542, 2017.
[33] Yan, R., Tang, J., Shu, X., Li, Z., & Tian, Q, "Participation-Contributed Temporal Dynamic Model for Group Activity Recognition," in *Proc. of the 26th ACM international conference on Multimedia*, 1292-1300, 2018. Article (CrossRef Link).

Bin Wang is a master student at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests involve Computer Vision and Pattern Recognition.

Wei Shen is a lecturer at the School of Communication and Information Engineering, Shanghai University, Shanghai, China and a full-time faculty at Department of Computer Science, Johns Hopkins University, Baltimore, USA. His research interests involve Computer Vision, Pattern Recognition and Machine Learning.

FanSheng Chen is an assistant researcher at Key Laboratory of Intelligent Infrared Perception , Chinese Academy of Sciences, Shanghai, China. His research interests involve High Spatial Resolution Remote Sensing, High Speed and Low Noise Information Acquisition Technology and Infrared Dim and Small Target Detection Technology.

Dan Zeng is a Full Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests involve Computer Vision, Multimedia Content Analysis, and Machine Learning.