Communication Papers of the Federated Conference on Computer Science and Information Systems pp. 129–134 DOI: 10.15439/2018F48 ISSN 2300-5963 ACSIS, Vol. 17

Soccer Object Motion Recognition based on 3D Convolutional Neural Networks

Jiwon Lee, Do-Won Nam, and Wonyoung Yoo
SW·Content Research Laboratory,
Electronics and Telecommunications Research Institute,
218 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea
Email: {ez1005, dwnam, zero2}@etri.re.kr

Yoonhyung Kim, Minki Jeong, and Changick Kim
Electrical Engineering,
Korea Advanced Institute of Science and Technology,
291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
Email: {yhkim, rhm033, changick}@kaist.ac.kr

***Abstract*—Due to the development of video understanding and big data analysis research field using deep learning technique, intelligent machines have replaced the tasks that people performed in the past in various fields such as traffic, surveillance, and security area. In the sports field, especially in soccer games, it is also attempting quantitative analysis of players and games through deep learning or big data analysis technique. However, because of the nature of soccer analysis, it is still difficult to make sophisticated automatic analysis due to technical limitations. In this paper, we propose a deep learning based motion recognition technique which is the basis of high level automatic soccer analysis. For sophisticated motion recognition, we maximize recognition accuracy by sequentially processing the data in three steps: data acquisition, data augmentation, and 3D CNN based motion classifier learning. As can be seen from the experimental results, the proposed method guarantees real-time speed performance and satisfactory accuracy performance.**

I. Introduction

IN the past, professional sports field was a human-oriented area. The training of the player has been done through the subjective guidance based on the know-how and experience of the manager and the coaching staff. Even in the case of a game judgement, it is judged through the intuition and observation of the referee, and the occasional misjudgement by the referee is accepted as part of sports. In addition, sports audiences were able to enjoy sports through unilateral delivery of sports contents. However, in recent years, many changes have been made in the field of professional sports as a result of quantitative analysis of sports through sports science and ICT technology. The manager and coaching staff can use data and video-based match analysis tools (eg, dartfish video analysis tool [1]) to check the objective player performance or conditions in detail, and to enable player training method or tactical changes. It also uses technology to help referee judges such as high-speed camera readings (eg, hawk-eye technology [2]) and produces interesting content using brilliant visualization tools (eg, freeD technology in NFL [3]) to give a sense of sports immersion. These sports analytic technologies are being developed to reflect the needs of people in many directions, thus the sports analysis market size was $4.7 Billions in 2017 [4].

This trend has also affected the professional soccer market. Germany World Cup is to take advantage of big data analytics company, SAP's Match Insights technology to improve the home team performance and analyze the strengths and weaknesses of the away teams to win the 2014 World Cup [5]. In addition to SAP, many international companies such as Chyronhego, OPTA, Deltatre, GPSports, and StatSport have technologies and services to perform quantitative analyzes on soccer matches and players.

In general, quantitative analysis of soccer game is consist of three steps: multi-object tracking, event analysis, and tactical analysis. Multi-object tracking can be automatically performed due to technological advances. However, in the cases of event analysis and tactical analysis, which require understanding of high level semantic from a given match, data is still extracted depending on the manual work of the expert group, and only the big data extracted by hand is secondarily processed and visualized. There are many reasons why these steps are not automated, but one of the biggest reasons is that the soccer event can be recognized only by the motion information of the player or referee. For example, it is necessary to be able to recognize a tackle motion of the player, a movement of the head referee's hand, and a flag motion of the assistant referee so that the tackle event, the foul event, and the offside event can be recognized. To solve this problem, this paper proposes a soccer object motion recognition technique.

This paper is composed as follows. In Sec. II, we describe the related researches. In Sec. III, we propose a soccer object motion recognition pipeline based on 3D convolutional neural networks (CNN). Sec. IV shows the experimental results of the proposed method. Finally, Sec. V discusses the concluding remarks.

II. Related Works

Motion recognition is a kind of computer vision field that recognizes human pose or action. The general process of motion recognition is as follows: 1) extracting feature points necessary for motion recognition in a given input source; 2) analyzing pattern of obtained feature points; 3) calculating similarity with predefined motion list; and 4) determining the final motion that has the highest similarity for the given input source. It is a kind of image classification technology in that the purpose of video-based motion recognition technology is to determine the final motion based on the similarity with predefined motion list.

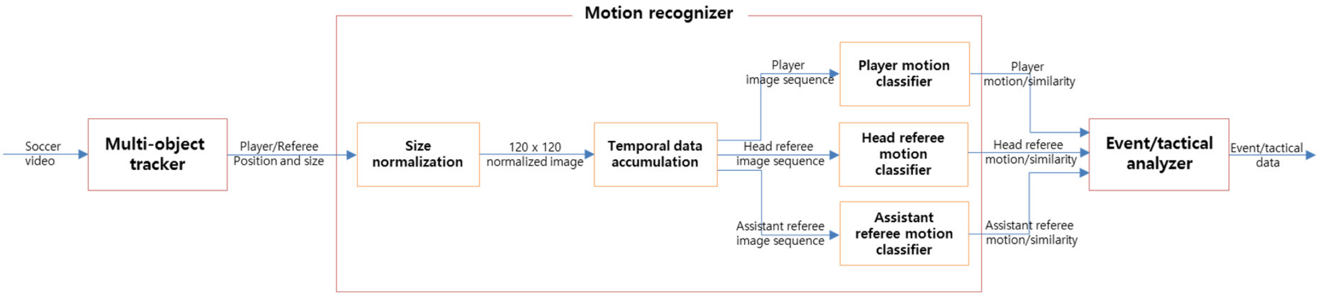

Motion recognizer
Soccer video
Multi-object tracker
Player/Referee Position and size
Size normalization
120 x 120 normalized image
Temporal data accumulation
Player image sequence
Head referee image sequence
Assistant referee image sequence
Player motion classifier
Head referee motion classifier
Assistant referee motion classifier
Player motion/similarity
Head referee motion/similarity
Assistant referee motion/similarity
Event/tactical analyzer
Event/tactical data

Fig. 1. System outline of the proposed method

Conventional motion recognition technologies are divided into several subdivisions according to several criteria. As a first criterion, it can be classified into two-dimensional(2D) motion recognition and three-dimensional(3D) motion recognition according to the dimension of the input source. 2D motion recognition technique performs motion recognition from 2D video sources taken from a general camera equipment [6], [7], [8]. 3D motion recognition technique performs motion recognition with stereoscopic video sources taken using special equipment such as MicroSoft's Kinect [9], [10]. As s second criterion, it can be classified as recognizing human action according to the human pose recognition and recognizing a gesture of a specific part of the human body. Motion recognition technique based on human pose tries to recognize motions such as human arm movements, arm extension, waist bending, and jumping motion based on video sources of human action [6], [7], [8], [9], [10]. Motion recognition technique for a specific gesture recognizes a partial movement of a specific part of the human body (hands, legs, *etc.*) [11], [12]. The third criterion is feature extraction method for motion recognition, and it is divided into hand-crafted feature extraction method and data-driven feature extraction method. A feature point is a clue that is used to distinguish different labels when performing motion classification. The accuracy of motion classification depends on the quality of the feature points. The hand-crafted feature extraction method is a method in which the user manually designs and extracts feature points according to a given classification purpose [6], [7], [8], [9], [10], [11], [12]. The hand-crafted feature extraction method are advantageous in that direct design of the user is easy and the patterns of motions to be classified are monotonous, but they have a disadvantage in that the performance is significantly lowered for motions with complex patterns. Recently, data-driven feature extraction method automatically learns feature points necessary for classification based on given information (video clip and label) [13]. Although this feature extraction method requires a large amount of computation and huge input data for learning, it performs much better than the hand-crafted feature extraction method in terms of accuracy and execution speed.

According to the above classification criteria, we can specify the category of motion recognition technique needed to solve the problem defined in this paper. In this paper, we use 2D video sources taken from camera equipment installed in the stadium. The target area of the field player and referees in the game is tracked, and the goal is to recognize the motion based on the tracking data. In addition, it is possible to construct large-sized learning data, which is suitable for data-driven feature learning and extraction. According to this analysis, the motion recognition technology proposed in this paper can be specified as 1) 2D video source based, 2) data-driven feature extraction, and 3) technology to recognize human pose.

III. Proposed Method

In this Section, a method of performing motion recognition by inputting object regions tracked from a soccer game video will be described in detail. Figure 1 depicts the system outline of the proposed method. For the motion recognition specialized for the soccer object, we constructed the motion recognition system through three steps of data acquisition, data processing, and motion classifier learning [14]. A detailed description of each is given in the subsection.

A. Data Acquisition

The data acquisition is performed first to recognize the motion. To do this, we need to define motion classification criteria. We classify motions of each soccer object and generate learning data based on the following principles:

- The object is categorized into field player, head referee and assistant referee.
- All the motions that each object can take on the field must be included in the motion list.
- The body direction of the object with respect to the same motion secures data of at least four directions.

TABLE I
Defined motion list for each soccer object

Field player	Head referee	Assistant referee
Stand		Sidle
Walk	Walk	Walk
Run in	Run	Run
Kick	One arm pointing	Flag up
Tackle/Lie	Card	Flag chest
Throw in		Flag side

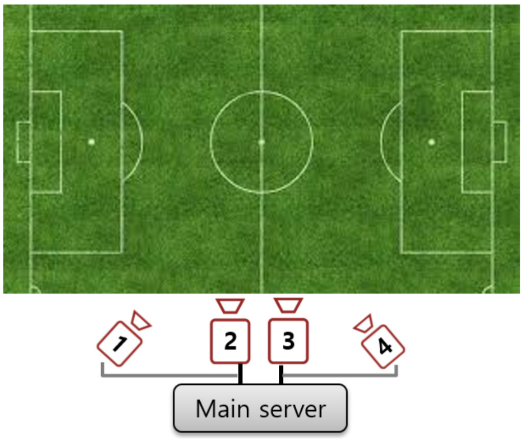


Fig. 2. Location of cameras in the stadium for data acquisition

- Motion with a duration of less than one second is excluded.

The motion classification list based on the above principles is shown in Table I.

After that, we acquired match videos through four cameras installed on a soccer stadium as shown in Fig. 2. In the figure, the first and fourth cameras shot the right half and the left half of the stadium respectively, and the second and third cameras shot the whole stadium. Here, the reason why the video was taken at various angles is to increase the recognition rate of the flag motion of the assistant referee. Then, the data necessary for motion classifier learning are acquired based on the object position extracted through the multi-object tracker [15].

B. Data Augmentation

After acquiring initial motion learning data, we extend the scale of learning data through data augmentation [16]. Motion data augmentation is closely related to the stability of the motion classifier. In general, the input to the motion classifier is a bounding box image including a soccer object which is an output of the multi-object tracker. Due to the nature of the tracking algorithm, the size and position of the bounding box fluctuate irregularly. Such trembling may cause performance degradation of motion recognition results. Therefore, a technique for effectively processing and extending motion learning data is needed for robust motion classifier that are robust to changes in bounding box size and position. We design a data processing algorithm suitable for this problem and incorporate it into motion classifier learning.

The designed data processing scheme is shown in Fig. 3, and its operation is as follows. First, a given image is normalized to an image of 140 pixels in width and in height. Then, a random cropping is performed at a size of 112 pixels in the horizontal and vertical directions. This process is introduced to imitate the phenomenon that the position of the bounding box shakes irregularly in the tracker output. Then, image up-scaling is performed in the following three ways for the image obtained through the random cropping. The first of the three ways is to use the original image, and the second and third are to perform up-scaling of the pre-processed image by 1.25 times and 1.5 times, respectively, and then random cropping (with a size of 112 pixels in the horizontal and vertical directions). As shown in Fig. 3, as a result of this process, the relative sizes of the objects existing in the pre-processing image are divided into three, and three images are obtained from one original image. This process is called scale augmentation, which imitates that the size of the bounding box changes irregularly in the tracker output. We have learned to diversify the relative size of the object to be recognized through the scale augmentation so that the classifier can be robust to the perspective of the tracked object. In order to secure the robustness against trembling phenomenon of the tracking result, randomly cropped data was learned through random cropping. The learning data that has been processed through data augmentation process is finally used as input to the motion classifier after normalization process with a size of 120×120 pixels.

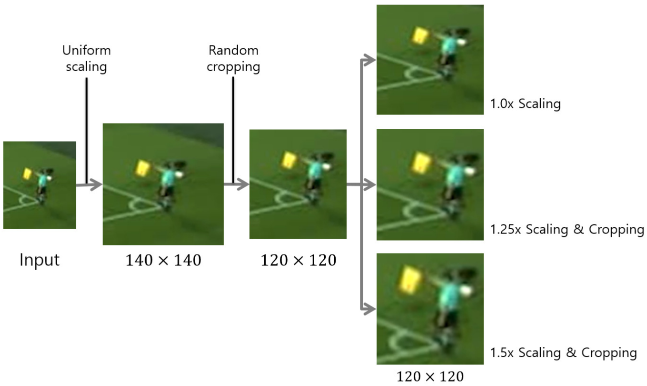


Fig. 3. Designed data augmentation scheme

C. Motion Classifier Learning

Finally, we have learned a deep learning based motion classifier based on acquired and augmented learning data. A motion classifier performs learning according to a predetermined number of labels at the training part and maps a given input image to one of the learned motion labels at the testing part. In this paper, we propose a 3D CNN-based motion classifier, which is a deep learning architecture that can understand the correlation between adjacent frames in order to take advantage of this feature, were used [17], [18].

Figure 4 shows a comparison of the 2D convolution and the 3D convolution. In the case of 2D convolution, a feature map

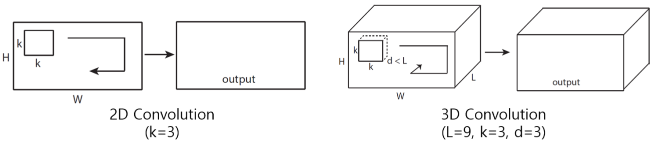


Fig. 4. Comparison of 2D convolution and 3D convolution

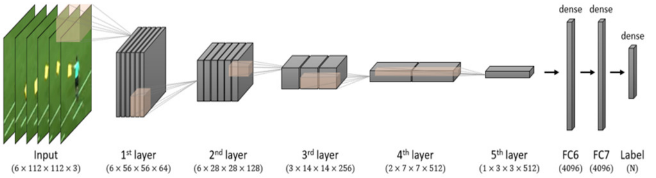


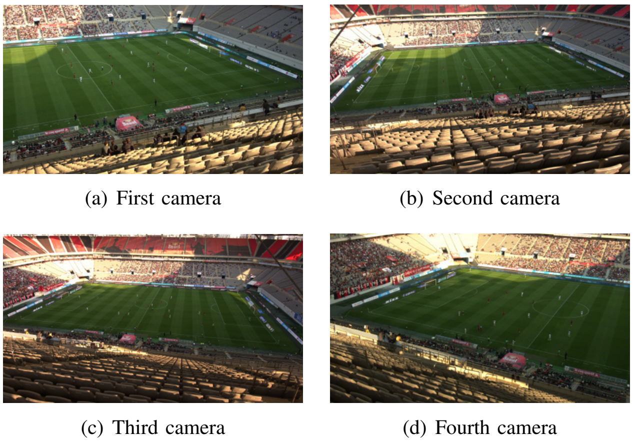
Fig. 5. Deep learning based motion classifier structure with 3D convolution (when N = 6)

is extracted using only spatial information for a single image, whereas a 3D convolution extracts not only spatial information but also temporal information for a plurality of continuous images to extract a feature map. Based on these features, the 3D CNN structure can be used to learn spatiotemporal information and contribute to performance enhancement and stabilization of the motion classifier. Figure 5 shows the network structure of a proposed motion classifier designed with a 3D CNN structure. The inputs to the network are F_b consecutive frame bundles (the frame bundle unit may vary depending on the applications), and a hierarchical feature map is extracted over a total of five layers for a given input video. In each feature map extraction step, 3D convolution is applied. A kernel having a differential depth (denoted by d in Fig. 5) is applied according to a layer of the feature map. The last three layers apply a fully connected network structure and apply a $softmax$ function to finally output the similarity for N motions. Here, the $softmax$ function is a generalization of the logistic function that normalizes a K-dimensional vector z having an arbitrary real values to a K-dimensional vector $\sigma(z)$ having real values in the range [0, 1] with a sum of 1. The function is given by

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, ..., K \tag{1}$$

IV. Experimental Results

In order to verify the performance of the proposed motion classifier, we took $4K$-sized videos at four different locations in the soccer stadium (see Fig. 2) for 9 K-league classic from 2016 to 2017. The captured video then proceeded to object tracking and divided the tracked data into field player, head referee, and assistant referee to generate learning data. A total of 170,000 pieces of initial learning data were generated, and about 600,000 pieces of final learning data were constructed after data augmentation process. Example screenshots of test videos and the generated learning data we have used are shown in Fig. 6 and Fig. 7, respectively.



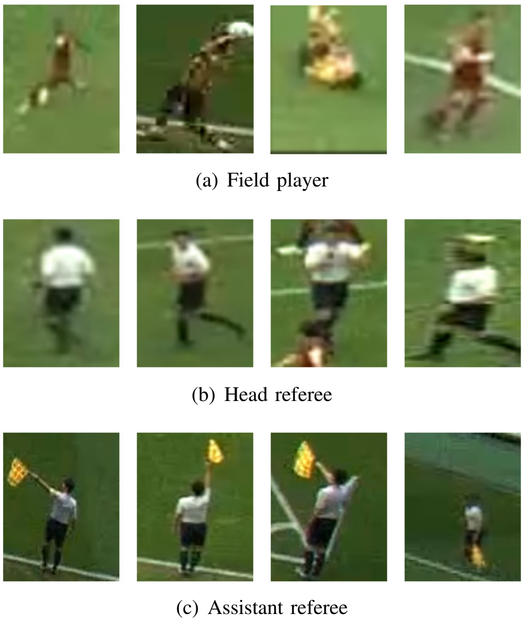
(a) First camera

(b) Second camera

(c) Third camera

(d) Fourth camera

Fig. 6. A sample screenshot of four cameras in the test video clip



(a) Field player

(b) Head referee

(c) Assistant referee

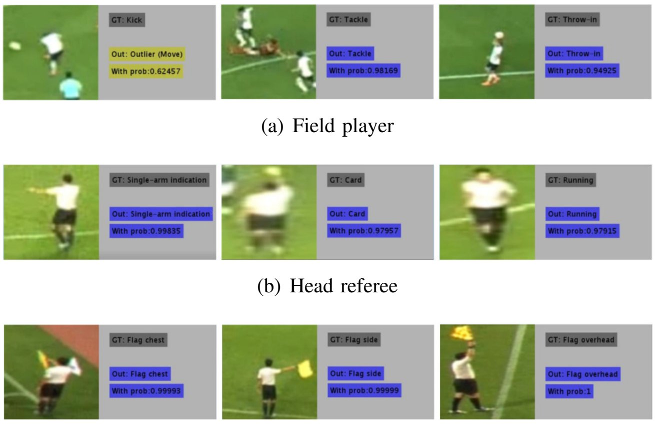
Fig. 7. An example of extracted data in each object

In the case of soccer game, 3D CNN based motion classifiers are designed to enable parallel processing using tensorflow [19], since the number of objects appearing in one frame is large at the same time. In order to improve the accuracy of motion classifier, we need to consider not only spatial clue but also temporal clue. Here, we provide different temporal clues according to the characteristics of object to be recognized. In more detail, since the motion of the field player occurs with a shorter duration than the motion of the referee, the field player classifies the motion into 4 frames by one unit($F_b = 4$), but in the case of the referee, 6 frames are grouped into one unit($F_b = 6$) to perform motion classification.

To evaluate the performance of the proposed motion classifier, we used the i7-6770 core processor, DDR3 64GB RAM, and three different GPUs. The performance of the motion

TABLE II
Performance variations of motion classifier for each GPU

GPU types	Performance	
	ops	fps
GeForce GTX 1070	800	32
GeForce GTX TITAN X	930	37.2
NVIDIA TITAN X	1200	48



(a) Field player

(b) Head referee

(c) Assistant referee

Fig. 8. Output examples of each motion classifier

classifier measured for each GPU is shown in Table II. Here, *ops* and *fps* refer to objects per second and frames per second, respectively. Since the number of field players and referees in one frame are 22 and 3, respectively, so the *fps* is calculated by dividing 25 from *ops* as shown in Table II. It means the proposed classifier has real-time motion recognition capability.

The finally obtained confusion matrix of each motion classifier is shown in Table III, IV, V, and the output example of the motion classifier is depicted in Fig. 8.

As can be seen from the experimental results, the average motion recognition accuracy of the field player, the head referee, and the assistant referee was 0.449, 0.851, and 0.872, respectively. It can be confirmed that the accuracy of the motion recognition of the field player is relatively low compared to the referee. In the case of the referees, it is easy to distinguish the motion because the number of motion to be recognized is small and the motion itself is stereotyped. However, in the case of the field player, the number of motions to be recognized is relatively large and the duration of occurred motion is also shorter than that of the referees (Recognition using only 66% frames compared to referees). In addition, the field player has a high degree of similarity between different motions, which is considered to have affected the accuracy.

TABLE III
CONFUSION MATRIX OF MOTION CLASSIFIER FOR FIELD PLAYER

Out\GT	**Stand**	**Walk**	**Run**	**Kick**	**Tackle /Lie**	**Throw in**
Stand	**1,779**	0	0	1	0	0
Walk	58	**325**	1,158	261	6	0
Run	0	0	**0**	4	0	0
Kick	569	694	702	**1,440**	1,147	648
Tackle/Lie	70	1	0	207	**768**	357
Throw in	9	254	37	308	35	**989**
Accuracy	**0.716**	**0.255**	**0**	**0.648**	**0.393**	**0.496**

TABLE IV
CONFUSION MATRIX OF MOTION CLASSIFIER FOR HEAD REFEREE

Out\GT	**Walk**	**Run**	**One arm pointing**	**Card**
Walk	**11,244**	202	376	17
Run	1,210	**10,579**	446	132
One arm pointing	395	499	**8,254**	143
Card	388	295	402	**546**
Accuracy	**0.850**	**0.914**	**0.871**	**0.652**

TABLE V
CONFUSION MATRIX OF MOTION CLASSIFIER FOR ASSISTANT REFEREE

Out\GT	**Sidle**	**Walk**	**Run**	**Flag up**	**Flag chest**	**Flag side**
Sidle	**12,610**	131	126	491	962	1,279
Walk	116	**12,564**	107	374	723	1,369
Run	51	159	**11,887**	216	430	342
Flag up	0	0	4	**9,033**	869	621
Flag chest	2	0	3	1,153	**16,030**	442
Flag side	46	18	5	1,616	1,317	**11,724**
Accuracy	**0.983**	**0.976**	**0.980**	**0.701**	**0.788**	**0.743**

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we introduced data acquisition, data processing, and motion classifier learning method to recognize motion of soccer objects from soccer video. In particular, to design motion classifiers with high accuracy, we use $3DCNN$, which is a structure that extracts spatio-temporal features well, and developed motion classifier considering real-time by using parallel processing technique. As can be seen from the experimental results, it can be seen that the proposed method satisfies the real-time speed performance and the high motion recognition accuracy of the referee. However, the accuracy of the recognition of field player motion is rather low, and further research is needed.

In the future, we will design a sophisticated motion classifier with high accuracy even for objects with ambiguous motion classification such as field player, and will try to incorporate the developed motion classifier into other sports fields such as basketball and figure skating.

ACKNOWLEDGMENT

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Reasearch & Development Program 2016 (R2016030044, Development of Context-Based Sport Video Analysis, Summarization, and Retrieval Technologies)

REFERENCES

[1] DartFish sports analysis tool [Online] Available : http://www.dartfish.com

[2] Hawk-eye innovations [Online] Available: https://www.hawkeyeinno vations.com

[3] FreeD on NFL [Online] Available : https://newsroom.intel.com/news/intel-nfl-kickoff-freed-technology-11-stadiums-create-immersive-highlights-2017-season/

[4] "Sports analytics: market shares, strategies, and forecasts, worldwide, 2015 to 2021," Wintergreen Research, 472 pages, May 2015

[5] A. Ghosh, "How 'Match Insight' is changing soccer," 6th Aug. 2014. [Online] Available: https://blogs.sap.com/2014/08/06/how-software- is-making-football-even-more-beautiful/

[6] C. P. Huang, C. H. Hsieh, K. T. Lai, and W. Y. Huang, "Human action recognition using histogram of oriented gradient of motion history image," *in International Conference on Instrumentation, Measurement, Computer, Communication and Control,* pp. 353-356, Oct. 2011.

[7] L. Hu, W. Liu, B. Li, and W. Xing, "Robust motion detection using histogram of oriented gradients for illumination variations," *in Proc. ICIMA 2010,* pp. 443-447, May. 2010.

[8] P. Banerjee and S. Sengupta, "Human motion detection and tracking for video surveillance," *in National Conference for Communication,* 2008.

[9] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," *in Proc. JCSSE 2012,* pp. 28-32, May. 2012.

[10] E. E. Stone and M. Skubic, "Fall detection in homes of older adults using the Microsoft Kinect," *IEEE Jour. Biomedical and Health Informatics,* vol. 19, no. 1, pp. 290-301, Mar. 2014.

[11] N. C. Kiliboz and U. Gudukbay, "A hand gesture recognition for human computer interaction," *Jour. Visual Communication and Image Representation,* vol. 28, pp. 97-104, Apr. 2015.

[12] M. B. Brahem, B. J. Menelas, and M. D. Otis, "Use of 3DOF accelerometer for foot tracking and gesture recognition in mobile HCI," *Peocedia Computer Science,* vol. 19, pp. 453-460, 2013.

[13] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *in Nature,* vol. 521, pp. 436-444, May. 2015.

[14] J. Lee, Y. Kim, M. Jeong, C. Kim, D. Nam, J. Lee, S. Moon, and W. Yoo, "3D convolutional neural networks for soccer object motion recognition," *in Proc. ICACT 2018,* pp. 354-358, Feb. 2018.

[15] W. Kim, S. Moon, J. Lee, D. Nam, and C. Jung , "Multiple Player Tracking in Soccer Videos : An Adaptive Multiscale Sampling Approach," *Multimedia Systems,* pp. 1-13, Feb. 2018.

[16] A. Krizhevsky, I. Sutskever, and G. E. hinton, "ImageNet classification with deep convolutional neural network," *in Proc. NIPS 2012,* pp. 1-9, Dec. 2012.

[17] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 35, no. 1, pp. 221-231, Mar. 2012.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *in Proc. ICCV 2015,* pp. 4489-4497, Dec. 2015.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv prepreprint arXiv:1603.04467v2,* 2016.