

3D Convolutional Neural Networks for Soccer Object Motion Recognition

Jiwon Lee*, Yoonhyung Kim**, Minki Jeong**, Changick Kim**, Do-Won Nam*, JungSoo Lee*,
Sungwon Moon*, and WonYoung Yoo*

* SW-Content Research Laboratory, ETRI(Electronics and Telecommunications Research Institute), Korea

**Electrical Engineering, KAIST(Korea Advanced Institute of Science and Technology), Korea

ez1005@etri.re.kr, yhkim1127@kaist.ac.kr, rhm033@kaist.ac.kr, changick@kaist.ac.kr, dwnam@etri.re.kr, jslee2365@etri.re.kr,
moonstarry@etri.re.kr, zero2@etri.re.kr

Abstract— Recently, sports and ICT technology have been combined, enabling quantitative and objective analysis of sports and athlete competence. In the case of soccer, quantitative analysis of competition and athletes is underway in various companies, but due to technical limitations, many data are still being generated based on the manual work of experts. In this paper, we propose an object motion recognition technique which is a basis for further automation of soccer analysis. We first classify objects in soccer game and define recognizable motion for each object category. After that, we design 3D CNN with spatiotemporal characteristics and extract the motion information that each object is currently taking from the match video. As can be seen from the experimental results, it can be confirmed that the proposed technique not only has higher speed performance than the existing methods, but also has high accuracy. In addition, it can be confirmed that there is a high possibility of expanding to areas such as CCTV surveillance.

Keywords— Motion recognition, Deep learning, 3D CNN, Soccer analysis, Sports science

I. INTRODUCTION

In most sports, the training of the athletes and the running of the games has been carried out through the guidance of professional coaching staff in the past. However, sports and ICT technologies have been combined to enable quantitative and objective analysis of the competences of sports and athletes recently. As a result, analysts in a number of sports fields have emerged, and the market for these sports analytics is expected to grow to 125 million dollars in 2014 and 4.7 billion dollars in 2017, based on a variety of devices and cloud services [1]. In the case of soccer, this tendency is followed by quantitative analysis of matches and players in international companies such as Deltatre, Chyronhego and OPTA.

In general, the quantitative analysis of soccer match and player consists of three stages of multi-object tracking, event analysis, and tactical analysis. Among these three stages, multi-object tracking can be automatically performed, but event analysis and tactical analysis stages still extract data depending on the manual operation of the expert group. There are many reasons why it is not possible to automate the stages after the multi-object tracking stage, but one of the biggest reasons is that the motion information of the player or the referee is needed to analysis the event. For example, to

recognize an offside event, we must be able to recognize the flag gesture of the assistant referee, the head referee's hand gesture to recognize the foul event, and the player's motion to recognize the tackle or heading event. To solve this problem, this paper introduces a soccer player and referee motion recognition technique as a basis for automating the quantitative analysis of a soccer match and players.

II. RELATED WORK

Motion recognition technique that identifies human behavior based on video plays a very important role in video based automation such as unmanned surveillance system and sports activity record system.

Until a while ago, motion recognition methods using handcrafted features such as Histogram-of-Oriented-Gradient (HOG) have been widely used, there has been a disadvantage that simultaneous motion recognition is impossible for many objects due to low processing speed and capacity [2]–[4]. In addition, if different image types with the trained type is inputted, it cannot work. Thus, the handcrafted features corresponding to the different types of images must be redesigned every time, so there is a limit to the range of application and accuracy. In order to solve the problems related to accuracy, motion recognition methods using equipment provided with object depth information have been proposed [5], [6]. These methods have cost disadvantages because they require additional special equipment such as Kinect. Therefore, there is a need for a motion recognition method that shows high accuracy and fast processing speed without special equipment.

Deep learning technology, on the other hand, is a representative example of data-driven feature extraction [7]. Since data-driven feature extraction requires a large amount of computation, actual utilization has not been high. However, in recent years, due to the enhancement of computer processing ability based on the improvement of semiconductor integration technology and the emergence of effective learning algorithms of data-driven features such as back propagation, the deep learning technique has developed at a high speed and it has been applied successfully in various fields such as image understanding, speech recognition.

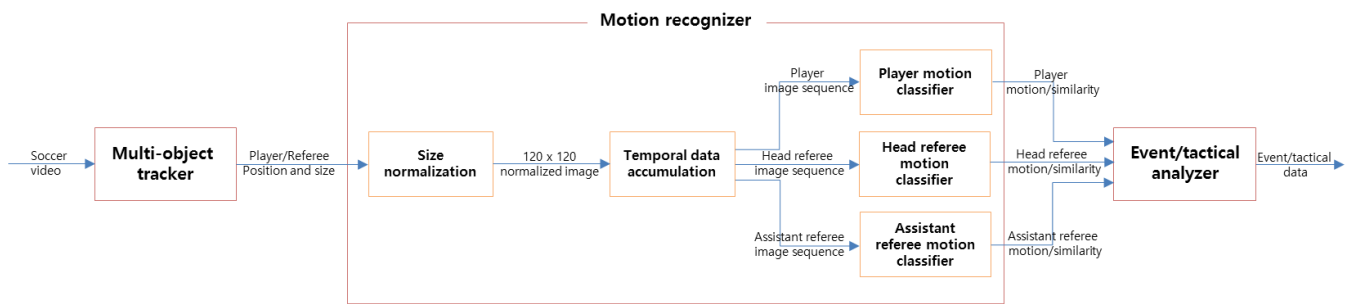


Figure 1. System architecture of the proposed method

III. PROPOSED METHOD

In this chapter, we introduce the method of motion recognition by inputting tracking result data from soccer video. Figure 1 shows the system architecture of the proposed method.

First, in order to perform motion recognition, it is necessary to classify the motions of the objects in the match. To do this, we classify motions and generate learning data based on the following principles:

- The object is divided into athlete, head referee and assistant referee.
- The motion category is set so that each object can include all the motions that it makes on the field.
- Exclude very short duration (less than 1 second) during motion.

The motions classified based on the above principle are shown in Table 1 below.

TABLE 1. DEFINED MOTION LIST FOR EACH OBJECT

Object type	Player	Head referee	Assistant referee
Motion list	Kick Tackle/Lie Throw in Stand Walk Run	One arm pointing Card Walk Run	Flag up Flag chest Flag side Walk Run Sidle

The motion learning data classified in this way is expanded through the data augmentation process. In general, a sufficient amount of learning data is needed to have generality of classifiers in a deep learning-based machine learning. Data augmentation is the answer to solve this problem. We obtained the robustness to the perspective by diversifying the relative size of the objects in the tracking result data through the scale augmentation. Through the random cropping, the classifiers learned about the irregularly cut data of the object from the multi-object tracker. The detailed data augmentation technique we performed is shown in fig. 2. The data that has completed the data augmentation is normalized to a size of 120×120 pixels for the training / test input data of the motion classifier in the next step.

Next, we trained deep learning based motion classifiers. Here, we have found that motion is meaningful when

considering consecutive frames in space-time axes. To take advantage of this feature, we used a motion classifier based on the 3D Convolutional Neural Network (CNN), a deep-running architecture that allows us to understand the correlation between adjacent frames [9], [10]. 3D CNN has a structure suitable to learn spatiotemporal tendency data unlike 2D CNN as shown in fig. 3, and it is easier to design parameters than Recurrent Neural Network (RNN). We trained the motion classifier with a series of continuous frame-based learning data predefined for each object class and then minimized the motion recognition time by conducting batch processing according to classification of each object when testing. Then, it is configured to provide the reliability of the output motion through the *softmax* function for the finally derived motion output. Here, the *softmax* function is a generalization of the logistic function that "squashes" a K -dimensional vector $\mathbf{z}$ of arbitrary real values to a K -dimensional vector $\sigma(\mathbf{z})$ of real values in the range $[0, 1]$ that add up to 1. The function is given by

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

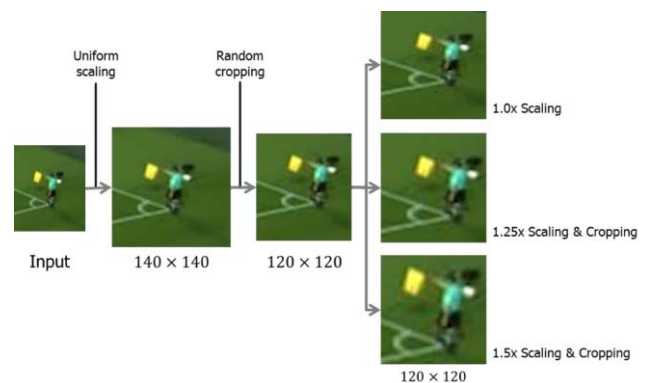


Figure 2. Data augmentation process

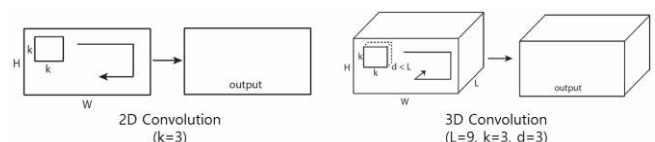


Figure 3. Difference between 2D convolution and 3D convolution

IV. EXPERIMENTAL RESULTS

In order to verify the performance of the proposed motion classifier, we selected 170 games of K-league classic matches from 2016 to 2017, taken at 4K size, and obtained 170,000 initial learning data. Through the data augmentation technique, we made a total of 600,000 learning data. An example of the captured video and extracted data is shown in fig. 4 and fig. 5.



Figure 4. A sample screenshot of test video clip



Figure 5. An example of extracted data in each object

The constructed learning data was used for training 3D CNN designed through *tensorflow* [11]. In order to provide spatiotemporal information to the motion classifier, 4 consecutive frames for the player and 6 consecutive frames for the head / assistant referee were grouped together in a single motion. Because the player motion occurs in a shorter time on the time axis than the head / assistant referee motion, the number of training frames of the player and the head / assistant referee are set differently to reflect this. The specification of the system used for motion recognition was intel i7 6770, 64GB memory and we performed experiments on three different GPUs. Table 2 shows the performance variations of motion test for each GPU.

TABLE 2. PERFORMANCE VARIATIONS OF MOTION TEST FOR EACH GPU

GPU series	Performance
NVIDIA GTX 1070 (Pascal arch.)	800 fps
NVIDIA TITAN X (Maxwell arch.)	930 fps
NVIDIA TITAN X (Pascal arch.)	1200 fps

In the table, 22 players recognize motion once every 4 frames, and every 3 head / assistant referees recognize motion once every 6 frames. Therefore, we can obtain the above

result when we refer to one frame. Using video frame based speed representation, the system shows a speed performance of about 45 fps on a TITAN X (pascal arch.) GPU when only motion recognition is performed. That is, the proposed technique has real-time motion recognition capability.

The finally obtained confusion matrix of each motion classifier is shown in Table 3-5, and the output example of the motion classifier is shown in fig. 6.

TABLE 3. CONFUSION MATRIX OF PLAYER MOTION CLASSIFIER

Out\GT	Kick	Tackle	Throw in	Stand	Walk	Run
Kick	1,440	1,147	648	569	694	702
Tackle	207	768	357	70	1	0
Throw in	308	35	989	9	254	37
Stand	1	0	0	1,779	0	0
Walk	261	6	0	58	325	1158
Run	4	0	0	0	0	0
Accuracy	0.648	0.393	0.496	0.716	0.255	0

TABLE 4. CONFUSION MATRIX OF HEAD REFEREE MOTION CLASSIFIER

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

TABLE 5. CONFUSION MATRIX OF ASSISTANT REFEREE MOTION CLASSIFIER

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

As can be seen from the experimental results, the accuracy of the head / assistant referee is high, but the accuracy of the player is quite low. In the case of the head / assistant referee, the number of motion to be recognized is small and the motion itself is stereotyped, so that it is easy to distinguish. In the case of the player, the number of motions to be classified is relatively large and the temporal clue is only 66% of that of the head / assistant referee, and the similarity between different motions is high. This is thought to have affected the accuracy.

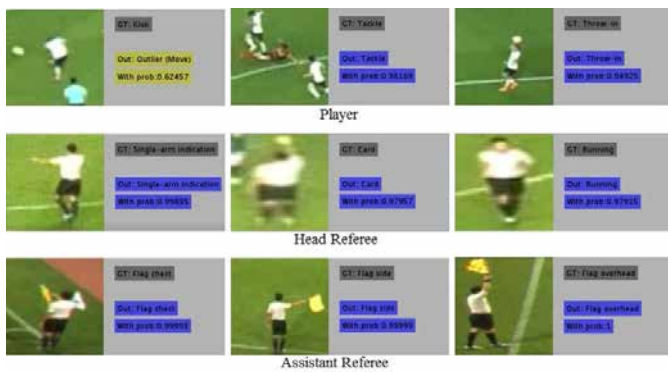


Figure 6. Output examples of each motion classifier

V. CONCLUSIONS

In this paper, we propose a method of automatically learning and extracting optimal feature information needed for motion recognition using 3D CNN, one of the deep learning technologies, and performing faster and more accurate motion recognition than conventional methods by applying a parallel processing algorithm. Respectively. As shown in the experimental results of the proposed technique, it guarantees real-time performance under given system specification and has high classification accuracy in the head / assistant referee motion classifiers. However, the player motion classifier has a somewhat low accuracy because of innate characteristics of player motions.

However, the proposed technique has advantages of fast motion recognition performance compared to existing technologies, and it is very useful in the field of image understanding applications where the market size is rapidly growing, such as multi-target unmanned surveillance system and sports player activity recording system.

In the future, we will reduce the categories of motions that need to be recognized by eliminating motions that are not directly relevant to the event recognition in the motion classification list, and we will expand the size of the learning data.

ACKNOWLEDGMENT

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

REFERENCES

- [1] "Sports analytics: market shares, strategies, and forecasts, worldwide, 2015 to 2021," Wintergreen Research, 472 pages, May 2015
- [2] 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.
- [3] 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.
- [4] P. Banerjee and S. Sengupta, "Human motion detection and tracking for video surveillance," in *National Conference for Communication*, 2008.
- [5] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," in *Proc. JCSSE 2012*, pp. 28-32, May. 2012.
- [6] 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.
- [7] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," in *Nature*, vol. 521, pp. 436-444, May. 2015.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural network," in *Proc. NIPS 2012*, pp. 1-9, Dec. 2012.
- [9] 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.
- [10] 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.
- [11] 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 preprint arXiv:1603.04467v2*, 2016.



image/video watermarking, and image/video processing.



Yoonhyung Kim received the B.S. and M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree at the School of Electrical Engineering, KAIST. His current research interests include image/video understanding, computer vision, and image processing.



Minki Jeong received the B.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2016. He is currently pursuing the Ph.D. degree at the School of Electrical Engineering, KAIST. His current research interests include image/video understanding, computer vision, pattern recognition, and 3D understanding.



Changick Kim received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 1989, the M.S. degree in electronics and electrical engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 1991, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 2000. From 2000 to 2005, he was a Senior Member of Technical Staff with Epson Research and Development, Inc., Palo Alto, CA, USA. From 2005 to 2009, he was an

Associate Professor with the School of Engineering, Information and Communications University, Daejeon, South Korea. Since March 2009, he has been with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, where he is currently a Professor. His research interests include multimedia signal processing, image/video understanding, and intelligent media processing.



Do-Won Nam received the B.S. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 1996 and M.S. degree in Information Technology from Pohang University of Science and Technology (POSTECH), Korea in 1998. He is working as a principal researcher in the Electronics and Telecommunications Research Institute (ETRI) since 2001. His research interests include data mining, digital rights management, digital cinema system and sports video analysis.



Jungsoo Lee received his B.S. and M.S. degrees from Jeonbuk University, Korea in 1995 and 1997, respectively and his Ph.D. degree in Electronic Engineering from Hanyang University, Seoul Korea in 2005. From 2000 to 2005, he was a senior member of MarkAny Research Institute. Currently, he is a senior member of Electronics and Telecommunications Research Institute (ETRI). His research interests are digital watermarking, fingerprinting, image

processing, digital rights management, digital cinema and digital signage.



Sungwon Moon received his B.S and M.S degrees of computer science from KAIST, Korea in 2010 and 2012. Since 2012 he has been a researcher in the Infocontent Section, SW.Content Reserch Laboratory, Electronics and Telecommunications Research Institute (ETRI), Republic of Korea. His research interests are Digital Watermarking, Video Forensic, and Video Processing.



Wonyoung Yoo received his Bachelor's, Master's and Ph. D. degrees in Electronic Engineering from Jeonbuk National University, Republic of Korea, in 1996, 1998 and 2003, respectively. Since 2001, he is working as a Managing Director of Electronics and Telecommunication Research Institute(ETRI). His research interests include image/video watermarking, fingerprinting and video analysis