Indonesian Journal of Electrical Engineering and Computer Science
Vol. 11, No. 3, September 2018, pp. 987~993
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v11.i3.pp987-993

A Survey of Video Based Action Recognition in Sports

Nur Azmina Rahmad[1], Muhammad Amir As'ari[2], Nurul Fathiah Ghazali[3], Norazman Shahar[4], Nur Anis Jasmin Sufri[5]
[2]Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[1,2,3,4,5]Sport Innovation and Technology Center (SITC), Institute of Human Centered Engineering (IHCE), Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Article Info

Article history:

Received Jan 25, 2018
Revised Apr 21, 2018
Accepted Jan 14, 2018

Keywords:

Action recognition
Deep learning
Sport performance analysis
Video based modality

ABSTRACT

Sport performance analysis which is crucial in sport practice is used to improve the performance of athletes during the games. Many studies and investigation have been done in detecting different movements of player for notational analysis using either sensor based or video based modality. Recently, vision based modality has become the research interest due to the vast development of video transmission online. There are tremendous experimental studies have been done using vision based modality in sport but only a few review study has been done previously. Hence, we provide a review study on the video based technique to recognize sport action toward establishing the automated notational analysis system. The paper will be organized into four parts. Firstly, we provide an overview of the current existing technologies of the video based sports intelligence systems. Secondly, we review the framework of action recognition in all fields before we further discuss the implementation of deep learning in vision based modality for sport actions. Finally, the paper summarizes the further trend and research direction in action recognition for sports using video approach. We believed that this review study would be very beneficial in providing a complete overview on video based action recognition in sports.

Corresponding Author:

Muhammad Amir As'ari,
Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia,
Johor Bahru, Malaysia.
Email: amir-asari@biomedical.utm.my

1. INTRODUCTION

Sports are synonym with the active movements of athletes on space either court or field. These movements are usually used by coach or trainer in evaluating the performance of their athletes. In sport practice, performance analysis can be divided into two: technique analysis and tactical analysis. According to Lees [1], technique analysis studies how the actions or movements were performed by the players. Tactical analysis or so called notational analysis studies what actions were carried out and the evaluation of the these actions take place [2]. Therefore, activity recognition is an important layer in tactical analysis before further analysis can be done by other researches. However, this paper will be focusing on the method in recognizing the action from sport video for establishing the automated notational analysis system.

Evaluating the players' performance has becomes a challenging task due to limitation on activity recognition phase. Hence, a technology intervention for games such as wearable sensor and video camera act as a tool to overcomes the challenges. Wearable sensor refers to the wearable device used by the athletes to collect the data of the activities in form of one-dimensional signals and the common wearable sensor used is inertial sensor [3]. Although this approach is effective to recognize the physical activities, unfortunately, the wearable sensor is less practical as athletes are not allowed to wear the sensors during the match. Not only that, the placement of sensors on to any part of human body also provide limitation to the movements of players during training or match.

Video camera is a visual sensing facility that provides visual data or video of the monitored activities and environmental changes. It has been used widely in high profile sports such as football and tennis for tactical analysis, tracking players, detecting the court lines, events recognition and video summarization [4-9]. The practicality of video based modality is more higher compared to the sensor based as no additional hardware will be attached to the athletes' body. Video camera captured the events and produced the broadcast video that receive high viewership. Therefore, these widely available video will be used by the researches to segment the useful part of video for the performance analysis [10].

Many research have implemented the video in recognizing the action, object or even vehicle [11-13]. But, the review study on this modality is still lacking. Therefore, this paper will be reviewed about the current and previous works on recognizing the action in sport using the video based modality.

2. VIDEO BASED SPORTS INTELLIGENCE TOOL

Since video based modality has attracted many researchers' attention in sport performance analysis, there are several tools developed by international company either for mobile or desktop usage. A summary of the existing intelligence system based on video for various sports is shown in Table 1.

Table 1. Present developed tool

Work	Description
[14]	A software developed by Dartfish company to highlight interested information from a match or training video for any sports. It is a useful for desktop and mobile tool for coach to analyses the performance of the athletes.
[15]	Vizrt introduced a few intelligence systems for modelling and animation, sport analysis, automated recording, graphic, video management and etc. for various high profile sports such as football, basketball, tennis and hockey.
[16]	Nacsport has been a marketer for video sport analysis software since 2008. It offers sport professionals to evaluate behaviors of all kinds of athletes. The data provided by software are both qualitative and quantitative and arranged according to need. Hence, the analysis can be done faster.
[17]	Sportradar provide services in collecting and analyzing sports data for over 1000 companies including sport federation, media companies and bookmakers.
[18]	Coach's eye is a video analysis app from TechSmith Corporation that analyses the video recorded from any device. It provides the real time analysis for individual athlete or team which enable coach to share the analyzed results immediately with the athletes for fast performance improvement.

However, in the aforementioned systems, there is no tool that can automatically recognize and classify the activities in sport for notational analysis. The highlighted actions for analysis are based on human perception. For example, in Dartfish, the analyst must manually watch, select and interpret the action of the athlete before further analysis by the software can be taken.

3. ACTION RECOGNITION FRAMEWORK

The video consists of temporal sequences of 2D images or can be defined as a set pixels in 3D space [19]. Figure 1 shows the overview of video based action recognition framework. According to [19-21], the action recognition can be divided into two approaches: simple approach and complex approach. Simple approach involves low level mechanism in which the recognized action is obtained from the detection and tracking of the human in each frame [22]. For example, to recognize "smash" activity in badminton match, firstly the players are detected from the surrounding background and then tracked to create a motion description of their movement. However, complex approach uses a lot of high level mechanism. This includes complex feature extraction and classification to recognize the human action.

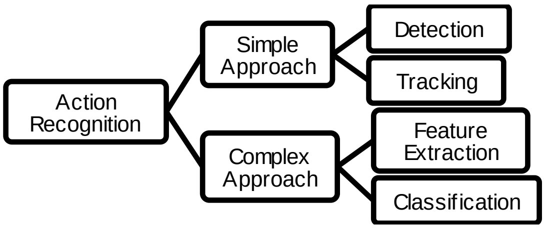

Figure 1. Action Recognition Approach

3.1. Detection and Tracking

In action recognition, tracking was used to generate the movement trajectory from the detected region of interest to infer the performed action. In [23], dense trajectories approach were proposed to describe videos. From each frame, dense points were sampled and tracked based on displacement information from a dense optical flow field. To evaluate video description in the context of action classification, a bag of features approach has been used. Proposed method shows an improvement on state-of-the-art on datasets with different level of difficulty such as KTH and Hollywood2. Since optical flow method is not suitable to be used in low resolution video, Zhao et.al [24] proposed a Region-based Mixture Model (RMM) for action classification of low resolution video. In this method, a set of long term motion trajectories and long term common shape is extracted from each video sequence using Layered Elastic Motion Tracking (LEMT) method. In addition, Particle Filter (PF) approach is a technique of Bayesian sequential importance sampling. Both work in [25, 26] utilized this approach in their studies. While work in [26] defined the human gestures and real time human tracking from depth data using PF method, work in [25] utilized a human-robot interface system which incorporates PF and Adaptive Multi-space Transformation (AMT) to track the pose of the human hand for controlling the robot manipulator. PF is used to estimate the translation of the human hand while AMT is used to improve the accuracy and reliability in determining the pose of the robot.

3.2. Feature Extraction and Classification

In machine learning, feature extraction is described as a pre-processing part to remove redundant part and reduce the dimensionality [27]. In [28], feature extraction was defined into low-level and high-level features. The key point for low level feature are corners, edges, blobs or contours while high level feature is more holistic like the structured information related to the action being taken. For sport video, important features including field, court, athlete and score board are extracted. However, classification is described as method to recognize the types of actions after feature extraction phase is completed.

In [29], different kind of features were extracted using fast feature descriptor which is 3D Histograms of Texture (3DHoTs) method. This method is derived from projecting depth frames onto frontal, side and top planes. And then, to classify the features a new classifier which is Multi-class Boosting Classifier (MCB) has been proposed. By providing a better margin distribution, the method was claimed to be efficient by maximizing the mean of margin whereas the variance of margin is still in minimum level. Not only that, work by Li et.al [30] presented an automatic players detection and analysing system in sports video sequences. There are three levels in recognizing the action of player in the moving background video. At granularity level, global motion estimation for filtering was proposed. Then, at the middle level, there is a segmentation of the highlighted object and finally at fine level, action recognition using Continuous Hidden Markov Model (CHMM) was proposed. As for work in [31], Hidden Markov Model (HMM) was trained to highlight the summary of RGB-D sport video that has been extracted by HAR method. Zhou et.al used linear Support Vector Machine (SVM) classifier to predict the class of action based on action representation after the extraction of local motion and appearance features has been done [32].

4. DEEP LEARNING IN VIDEO BASED APPROACH

Deep learning is a subtype of machine learning but more promising approach as compare to other conventional machine learning approaches. Deep learning is similar with one of the machine learning model called artificial neural network layer. Both consist of input layer, hidden layer and output layer. But, the number of hidden layer in deep learning could reach to hundreds layer and this is where the term “deep” is came from. Since deep learning shows an impressive result on several applications such as image classification, it has been implemented in the action recognition application. Table 2 shows the differences between deep learning and machine learning in term of preprocessing phase, size of data set, training time and hardware requirement. In deep learning, preprocessing phase is not required. As illustrated in Figure 2, deep learning eliminates the manual feature extraction phase because the network extract features directly from images during training. To train the deep learning model, large data sets are required to compensate large size of hidden layers. However, due to the vast development of broadcast sport video online which is accessible, sport video analysis using deep learning has become the emerging research interest [33]. Due to complexity of the deep learning, more time is needed to train the model. Hence, high performance GPU is important to reduce the training time. GPU is chosen compared to CPU because it has parallel architecture that accelerates the computing process. One of the most popular deep learning model is Convolutional Neural Network (CNN). Table 3 shows a few various types of CNN which were formulated to classify difference type of action.

Table 2. Machine Learning vs. Deep Learning

Characteristic	Machine learning	Deep learning
Preprocessing phase	Need	Does not need
Size of data set	Small	Large
Training time	Short	Long
Hardware requirement	Simple	High end

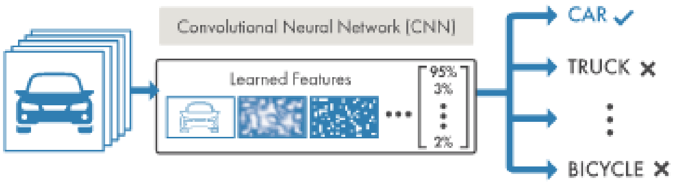

Figure 2. Illustration of deep learning architecture [34]

Table 3. Summary of Types of CNN

Work	Method
[35]	Deep ConvNet
[36]	3D ConvNet
[37, 38]	Two-stream ConvNet

The work in [35] modelled the Convolutional Neural Network (CNN) to classify 1 million Youtube videos contain 487 classes (called Sports-1M dataset). The model which can be seen in Figure 3 was divided to process the input into low resolution context stream and high resolution fovea stream before the networks was trained using UCF-101 dataset alone in order to increase the performance of runtime. Both streams consist of convolutional, normalization and pooling layers that alternate each other and the two streams finally converged into two fully connected layers. The performance of top layers on UCF-101 dataset show significant improvement compared to the UCF-101 baseline model.

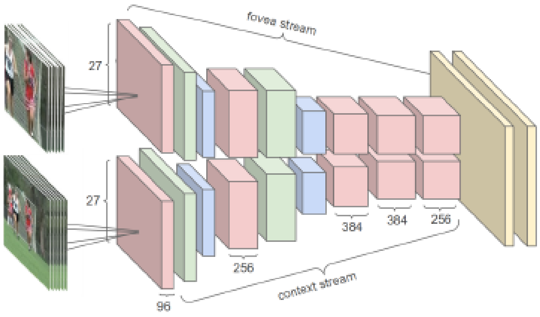

Figure 3. CNN architecture [35]

One of the major issues in action recognition is high inter and intra class variations and large class imbalance. Hence, to overcome the aforementioned problem, work in [36] implemented a 3D Convolutional Neural Network (CNN) for multi-label class-imbalanced in hockey videos on two deep approaches: 1) ensemble of k-binary network; and 2) single multi-label k-output network. The 3D convolutional and pooling process were embedded in the proposed approaches to tackle the multi label recognition.

Wang et.al [37] proposed Trajectory-Pooled Deep-Convolutional Descriptors (TDD) conducted on two challenging datasets: HMDB51 and UCF101. In this method, firstly, the two-stream ConvNets were trained on both datasets as a deep ConvNet to extract multiclass feature maps from video sequences [38]. Then, the TDD descriptor was obtained using pooling process of these ConvNets before classifying the action using SVM classifier to perform action recognition.

Besides CNN, another model of supervised deep learning which is good in handling sequential data is recurrent neural network (RNN) and one of the common RNN is called Long Short-Term Memory

(LSTM) [39, 40]. The work in [36] which explained earlier also attempt to embedded the proposed 3D CNN with LSTM. The LSTM was added in between flattened CNN feature layers and dense layers (see Figure 4). However, the result shows that the combination of LSTM reduces the performance because the short sequences of frame were used.

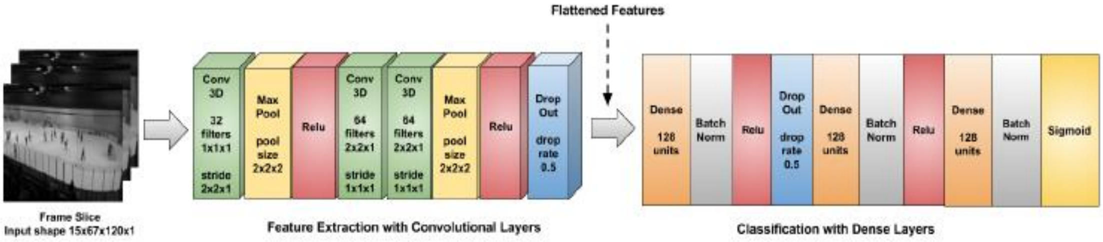

Figure 4. General structure of the network [36]

Not only that, there are several attempts to implement LSTM in football video analysis. For example in [41], LSTM was proposed to classify football video sequence on MICC-Soccer-Actions-4 dataset which contains four action classes using visual and motion content. In this study, the features that represent the visual content were established using the Bag of Words (BoW) technique while SIFT-based model was proposed to extract the motion features. These features were used in LSTM to classify the action. The results show that the performance in action classification for the combination proposed models is 92%. Tsunoda et.al [42] also worked on action recognition for football video by implementing hierarchical LSTM. In the proposed model, several CNN models were integrated with two layers of LSTMs (see Figure 5). CNN was used to extract multiple person-centered features. Then, the first layer of LSTM was computed to integrate all k-numbers of person-centered features before the last LSTM layer integrated the temporal sequence of integrated multiple person features. The work in [43] focused in implementation of LSTM in ice hockey video which has been extracted into sequence of images to classify five puck possession events: dump in, dump out, pass, shot and loose puck delivery. In this work, firstly, the whole frame and each player were extracted using pre-trained CNN to obtain the content information, individual action and interaction between players. The pre-trained AlexNet model was chosen for the extraction phase as the number of available data is small and the model showed the great achievement in various computer vision tasks. Later in events prediction phase, one layer of LSTM model is used to classify the five puck possession events.

However, in [44] , deep fusion framework was introduced by combining spatial features from CNN with temporal features from LSTM on three datasets: UCF11, UCFSports and jHMDB. Four CNNs and LSTM fusion methods for the recognition of human actions were proposed (two direct mapping models and two merged models). First two models are single stream models called as conv-L and fc-L (direct mapping models). These models extract CNN activation outputs from the last convolutional layer and the first fully connected layer for each frame of each video. The final output of these models is determined by considering the output from the soft-max layer of LSTM network which fed with features obtained from the CNN. The second remaining models called fu-1 and fu-2 (two merged models) are two stream approaches where two networks are merged. From the experiment, it shows that the direct mapping methods are less accurate compared to the two merged models. But, among two merged models, fu-2 shows the good accuracy value. It proved that this fusion method produces best results with the aid of deep layer wise structure.

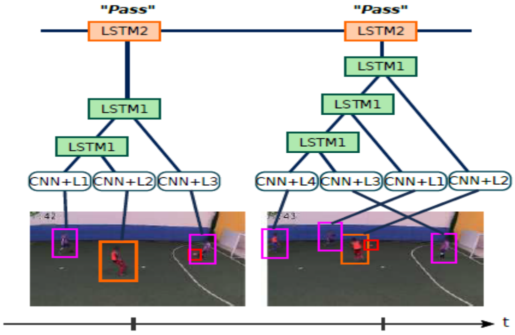

Figure 5. Hierarchical LSTM [42]

5. CONCLUSION AND FURTHER RESEARCH DIRECTION

Due to the rapid development of broadcast video online on sport match, it has become a tool for action recognition for sport analysis. Two main approaches used by researches are machine learning and deep learning. There are various works have been done on both approaches. Recently, deep learning approach such as CNN and RNN have been tremendously used in many works as it provides a better accuracy and capable of eliminating the complex preprocessing phase. However, it has becoming the issues since each proposed method can only classify the actions for certain sport because different sport has different context and features. Hence, in future, a flexible method for action recognition can be proposed in which one method can classify different type of sports.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Universiti Teknologi Malaysia (UTM) for supporting this research under Zamalah Scholarship.

REFERENCES

[1] Lees A. “Technique analysis in sports: A critical review”. *Sports Science.* 2002; 20.(10): 813-828.

[2] Hughes M, Bartlett R.M. “The use of performance indicators in performance analysis”. *Sports Science.* 2002; 20.(10):739-754.

[3] Dominguez Veiga J.J, O'Reilly M, Whelan D, Caulfield B, Ward E.T. “Feature-Free Activity Classification of Inertial Sensor Data With Machine Vision Techniques: Method, Development, and Evaluation”. *JMIR mHealth and uHealth.* 2017; 5.(8): 115.

[4] Choroś K. “Detection of Tennis Court Lines for Sport Video Categorization”. *Computational Collective Intelligence Technologies and Applications: 4th International Conference.* Ho Chi Minh City, Vietnam. 2012: 304-314.

[5] Kapela R, Świetlicka A, Rybarczyk A, Kolanowski K, O′Connor N.E. “Real-time event classification in field sport videos”. *Signal Processing: Image Communication.* 2015; 35.(1): 35-45.

[6] Lai J.-H, Chen C.-L, Kao C.-C., Chien S.-Y. “Tennis Video 2.0: A new presentation of sports videos with content separation and rendering”. *Journal of Visual Communication and Image Representation.* 2011; 22.(3): 271-283.

[7] Niu Z, Gao X, Tian Q. “Tactic analysis based on real-world ball trajectory in soccer video”. *Pattern Recognition.* 2012; 45.(5): 1937-1947.

[8] Conaire C, Kelly P, Connaghan D, O'Connor N.E. “TennisSense: A platform for extracting semantic information from multi-camera tennis data”. *Digital Signal Processing, 2009 16th International Conference.* Santorini-Hellas, Greece. 2009: 1-6.

[9] Sun L, Liu G. “Field lines and players detection and recognition in soccer video”. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing.* Taipei, Taiwan. 2009: 1237-1240.

[10] Zhu G, Xu C, Huang Q, Rui Y, Jiang S, Gao W, Yao H. “Event Tactic Analysis Based on Broadcast Sports Video”. *IEEE Transactions on Multimedia.* 2009; 11.(1): 49-67.

[11] Lu A, Zhong L, Li L, Wang Q. “Moving Vehicle Recognition and Feature Extraction From Tunnel Monitoring Videos”. T*ELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11.(10): 6060-6067.

[12] Wang L, Yun T, Lin H. “Boost Action Recognition through Computed Volume”. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11.(4): 1871-1876.

[13] Xu P, Qingdao U. “Study on Moving Objects by Video Monitoring System of Recognition and Tracing Scheme”. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11.(9): 4847-4854.

[14] Dartfish Website. [Online]; Available from: http://www.dartfish.com/360_S.

[15] Virtz Website. [Online]; Available from: http://www.vizrt.com.

[16] NacSport Website. [Online]; Available from: http://www.nacsport.com.

[17] SportRadar Website. [Online]; Available from: http://sportradar.com.

[18] Coach's Eye Sports Video Analysis App. [Online]; Available from: https://www.coachseye.com/.

[19] Cheng G, Wan Y, Saudagar A.N, Namuduri K, Buckles B.P. “Advances in Human Action Recognition: A Survey”. *CoRR.* 2015; 1501.05964.(1): 1-30.

[20] Xu X, Tang J, Zhang X, Liu X, Zhang H, Qiu Y. “Exploring Techniques for Vision Based Human Activity Recognition: Methods, Systems, and Evaluation”. *Sensors.* 2013; 13.(2): 1635-1650.

[21] Poppe R. “A Survey on Vision-based Human Action Recognition”. *Image and Vision Computing 2010.* 2010; 28.(6): 976-990.

[22] Aggarwal J.K, Ryoo M.S. “Human activity analysis: A review”. *Journal ACM Computing Surveys (CSUR).* 2011; 43.(3).

[23] Wang H, Kläser A, Schmid C, Liu C.L. “Action recognition by dense trajectories. Computer Vision and Pattern Recognition (CVPR)”, *2011 IEEE Conference. Providence*, RI, USA. 2011: 3169-3176.

[24] Zhao Y, Di H, Zhang J, Lu Y, Lv F. “Recognizing human actions from low-resolution videos by region-based mixture models”. *2016 IEEE International Conference on Multimedia and Expo (ICME).* Seattle, WA, USA. 2016: 1-6.

[25] Du G, Zhang P, Wang X. "Human-Manipulator Interface Using Particle Filter". *The Scientific World Journal*. 2014.(2014): 1-12.

[26] Bednaˇr´ık J, Herman D. "Human gesture recognition using top view depth data obtained from Kinect sensor". *The Excel @ FIT Conference. Faculty of Information Technology of the Brno University of Technology*. 2015.

[27] Khalid S, Khalil T, Nasreen S. "A survey of feature selection and feature extraction techniques in machine learning". *2014 Science and Information Conference.* London, UK. 2014:372-378.

[28] Soomro K, Zamir A.R. "Action Recognition in Realistic Sports Videos". USA: Springer. 2014.

[29] Zhang B, Yang Y, Chen C, Yang L, Han J, Shao L. "Action Recognition Using 3D Histograms of Texture and A Multi-Class Boosting Classifier". *IEEE Transactions on Image Processing*. 2017; 26.(10): 4648-4660.

[30] Li H, Tang J, Wu S, Zhang Y, Li S. "Automatic Detection and Analysis of Player Action in Moving Background Sports Video Sequences". *IEEE Transactions on Circuits and Systems for Video Technology.* 2010; 20.(3): 351-364.

[31] Tejero-de-Pablos A, Nakashima Y, Sato T, Yokoya N. "Human action recognition-based video summarization for RGB-D personal sports video". *2016 IEEE International Conference on Multimedia and Expo (ICME)*. Seattle, WA, USA. 2016: 1-6.

[32] Zhou W, Wang C, Xiao B, Zhang Z. "Action recognition via structured codebook construction". *Signal Processing: Image Communication*. 2014; 29.(4): 546-555.

[33] Rahmani H, Mian A, Shah M. "Learning a Deep Model for Human Action Recognition from Novel Viewpoints". *CoRR*. 2016; abs/1602.00828.(1): 1-14.

[34] Lautenbach F. "A laboratory study on attentional bias as an underlying mechanism affecting the link between cortisol and performance, leading to a discussion on the nature of the stressor (artificial vs. psychosocial)". *Physiology & Behavior*. 2017; 175.(175): 9-15.

[35] Karpathy A, Toderici G, Shetty S, Fei L.F. "Large-Scale Video Classification with Convolutional Neural Networks". *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Columbus, OH, USA. 2014:1725-1732.

[36] Sozykin K, Khan A.M, Protasov S, Hussain R. "Multi-label Class-imbalanced Action Recognition in Hockey Videos via 3D Convolutional Neural Networks". *Computer Vision and Pattern Recognition*. 2017; abs/1709.01421.(1): 1-8.

[37] Wang L, Qiao Y, Tang X. "Action recognition with trajectory-pooled deep-convolutional descriptors.2015 IEEE" *Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA. 2015: 4305-4314 .

[38] Simonyan K, Zisserman A. "Two-Stream Convolutional Networks for Action Recognition in Videos". *Neural Information Processing Systems Conference*. Montreal,Canada. 2014: 1-11.

[39] Gers F.A, Schraudolph N.N, Schmidhuber J. "Learning precise timing with LSTM recurrent networks". *The Journal of Machine Learning Research*. 2003; 3.(1): 115-143.

[40] Hochreiter S, Schmidhuber J. "Long Short-Term Memory". *Journal Neural Computation*. 1997; 9.(8): 1735-1780.

[41] Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A. "Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks". *Artificial Neural Networks – ICANN 2010: 20th International Conference*. Thessaloniki, Greece. 2010: 154-159.

[42] Tsunoda T, Komori Y, Matsugu M, Harada T. "Football Action Recognition Using Hierarchical LSTM. 2017". *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* Honolulu, HI, USA. 2017: 1-9.

[43] Tora M.R, Chen J, Little J. J. "Classification of Puck Possession Events in Ice Hockey". *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Honolulu, HI, USA. 2017: 1-8.

[44] Gammulle H, Denman S, Sridharan S, Fookes C. "Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition". *CoRR*. 2017; abs/1704.01194.(1): 1-10.