Vision Based Automated Badminton Action Recognition Using the New Local Convolutional Neural Network Extractor

Chapter · April 2020

DOI: 10.1007/978-981-15-3270-2_30

CITATION
1

READS
265

5 authors, including:

Nur Azmina Rahmad
Universiti Teknologi Malaysia
11 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Muhammad Amir As'ari
Universiti Teknologi Malaysia
31 PUBLICATIONS **98** CITATIONS
SEE PROFILE

Nur Anis Jasmin Sufri
Universiti Teknologi Malaysia
7 PUBLICATIONS **37** CITATIONS
SEE PROFILE

Keerthana Rangasamy
Universiti Teknologi Malaysia
3 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Vision Based Automated Badminton Action Recognition Using the New Local Convolutional Neural Network Extractor

Nur Azmina Rahmad[1], Muhammad Amir As'ari[2(✉)], Mohamad Fauzi Ibrahim[3], Nur Anis Jasmin Sufri[1], and Keerthana Rangasamy[1]

[1] School of Biomedical Engineering and Health Sciences, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[2] Sport Innovation and Technology Center (SITC), Institute of Human Centered Engineering (IHCE), Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
amir-asari@biomedical.utm.my
[3] Institut Sukan Negara, Kompleks Sukan Negara, Bukit Jalil, 57000 Kuala Lumpur, Malaysia

Abstract. Performance analysis is essential in sports practice where the athlete is evaluated to improve their performance. Due to the rapid growth of science and technology, research on automated recognition of sports actions has become ubiquitous. The implementation of automated action recognition is an effort to overcome the manual action recognition in sport performance analysis. In this study, we developed a model for automated badminton action recognition from the computer vision data inputs using the deep learning pre-trained AlexNet Convolutional Neural Network (CNN) for features extraction and classify the features using supervised machine learning method which is linear Support-Vector Machine (SVM). The data inputs consist of badminton match images of two classes: hit and non-hit action. Before pre-trained AlexNet CNN was directly extracting the features, we introduced the new local CNN extractor in recognition pipeline. The results show that the classification accuracy with this new local CNN method achieved 98.7%. In conclusion, this new local CNN extractor can contribute to the improvement of the performance accuracy of the classification task.

Keywords: Action recognition · Deep learning · Convolutional Neural Network · Badminton

1 Introduction

Nowadays, sport performance analysis is not something new in the sport field. It has been widely used by coaches and analyst experts to evaluate and improve the performance of athletes during the competition, sport event or coaching session. Due to a massive access to technology and computer science, the analysis can be done easily

M. H. A. Hassan et al. (Eds.): MoHE 2019, LNBE, pp. 290–298, 2020.
https://doi.org/10.1007/978-981-15-3270-2_30

using the vision based modality on the established performance analysis software such as Dartfish, Nacsport, Performasports and LongoMatch. However, the major issue of the current performance analysis is manual action annotation that need to be done by analyst is unpractical, time consuming and susceptible to human error. Therefore, studies to develop an automated human action recognition (HAR) in sports have been proposed by many researchers by using either conventional machine learning or deep learning approach. The establishment of automated action recognition is beneficial for coaches and analyst experts to make their analysis more effective and accurate [1]. In HAR task, there are several techniques introduced by previous researchers. The first one is handcrafted features extraction and classification using conventional machine learning technique such as decision tree, support vector machine, hidden markov models and naïve Bayes [2, 3]. Secondly, the most promising and popular technique nowadays is deep learning technique [4].

In deep learning, Convolutional Neural Network (CNN) is the common approach that has been used for HAR from the video frame that represent human action [5–7]. This is due to the fact that CNN is excellent in image recognition task which is almost similar to video frame recognition task in video based recognition problem. However, the conventional CNN that is applied directly and globally to the whole image frame pixels is not susceptible to the influence of background pixels which might fail the recognition process especially when the region of interest in the input video frame is too small. Thus, the focus of this paper is to discuss the proposed automated badminton action recognition (between hit and non-hit action) from the broadcast video of badminton match using new proposed local CNN extractor and linear SVM classifier which outperform the conventional global CNN approach.

2 Related Works

Human action recognition (HAR) is a task of formulating an automated algorithm for recognizing or classifying human actions which in general can be divided into two categories: (1) HAR based on wearable sensor; (2) HAR based on vision sensor. HAR involves in many applications such as video surveillance [8], human-computer interaction, virtual reality systems, human monitoring [9–12] and sport analytic systems [13, 14]. Wearable sensor refers to the method of positioning the sensor such as inertial sensor and accelerometer [15–17] at human body that produce one-dimensional signals related to the body movement. Meanwhile, vision sensor is the approach that focused on the special camera setup or Kinect [18] to capture the information in form of video sequences. HAR is a challenging task for both modalities in which sensor based faces problems such as lack of information as it depends on one-dimensional signals and less practical because it bounded on the multiple sensors system for more accurate recognition. As for vision based, the challenges are occlusion, illumination and view point variations and background differences [19] on the video data inputs.

In HAR based on vision sensor task, there are several techniques introduced by previous researchers. The first one is handcrafted features extraction and classification using conventional machine learning technique such as decision tree, support vector machine, hidden markov models and naïve Bayes [2, 3]. Previously, most of the works

were done using the handcrafted machine learning approach such as works in [2, 20–22]. Secondly, the most promising and popular technique nowadays is deep learning technique [4, 23]. The use of deep learning methods in recognizing human actions is not new in computer vision field, as many researches have been done in this area. Due to the availability of enormous dataset, deep learning technique has becoming a bombing interest in action recognition especially using CNN method. The current machine learning technique is relying on the heuristic handcrafted features extraction which is bounded to the human domain knowledge. In brief, features extraction need to be done manually before classifiers classify the shallow learned features [24]. Deep learning technique is subtype of machine learning. It is an improved technique that eliminates the manual features extraction in machine learning pipeline. The features are being learned automatically from the low-level features at the first layer until high-level features at the deepest layer to perform classification task directly from the input data. Figures 1 and 2 illustrate the machine learning and deep learning pipeline. There are various methods in deep learning such as deep stacking network (DSN), deep belief network (DBN), recurrent neural network (RNN), long-short term memory (LSTM) and frequently used in many research is CNN.

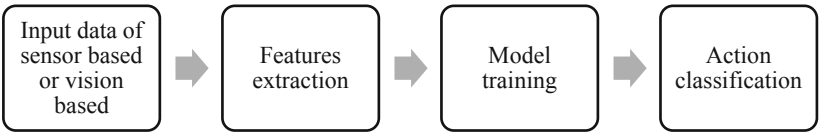

Fig. 1. Machine learning pipeline

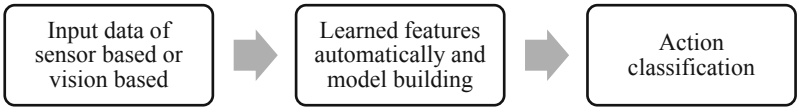

Fig. 2. Deep learning pipeline

As a deep learning network, CNN comprises of input layer, hundreds of hidden layers and output layer. The features were attained from the convolutional process between filters and input data. Next, the activated features were transmitted from one layer to the next layer before being classified at the final fully connected layers. There are mainly three methods to use CNN which depending on the type of application – training from scratch, transfer learning of pre-trained CNN or feature extraction. Training deep learning from scratch required a huge number of labelled dataset and longer training period. This approach however is suitable to be used for a new application or application with many output classes. Second method is transfer learning approach which researchers mostly use. This method involves fine-tuning the pre-trained model such as AlexNet [25], GoogleNet, SqueezeNet, ResNet and VggNet [7] with own labelled dataset. After tuning the model with own dataset, a new classification task can be performed. The advantage of this method compared to training from scratch method is that it required less number of dataset. Hence, the training time is faster.

Lastly, it can also be used as more specialized approach which is feature extractor. Since features were learned in each layer, features can be extracted to be classified by other machine learning model such as SVM.

Some researchers reviewed sport action recognition based on deep learning methods and generally, most are focusing on football and basketball [26]. Due to the rapid development of technology nowadays, most of the researchers are utilizing the video-content analysis approach for sport action recognition. For example, work in [27] used video dataset to classify five hockey actions using pre-trained CNN method and [5] classify low and high resolution video dataset using CNN method. Meanwhile, Baccouche *et al.* [28] used subtype of Recurrent Neural Networks (RNN) which is LSTM to recognize soccer actions. In deep learning, the method that is widely used is CNN. However, in recognizing different type of actions with excellent performance, researchers have proposed the method of fusing few deep learning methods. For example, work in [29] used video dataset to recognize sport actions using CNN and bi-directional Long Short Term Memory (LSTM) deep learning method.

As in this paper, we utilized the pre-trained AlexNet CNN model as a features extractor and SVM as a classifier to classify hit and non-hit actions from badminton broadcast videos. We also proposed our new local CNN extractor method in deep learning pipeline to improve the performance accuracy of the deep learning model in classifying the badminton actions.

3 Methodology

In this section, the proposed method and the experimental framework are discussed in detail. The experiment was conducted on our own constructed dataset obtained from five broadcasted badminton match videos. There are two classes of dataset which is hit and non-hit action of total 2990 images compressed to 227*227 where 60% from the total images sample are used for training, 20% for testing and another 20% for validation. The details of dataset used in this experiment are as follow (Table 1).

Table 1. Details of dataset

Class	Number of training sample	Number of testing sample	Number of validation sample
Hit	897	299	**299**
Non-hit	897	299	**299**
Total	**1794**	**598**	**598**

Figure 3 shows the methodology of the experiment. Unlike the normal deep learning pipeline that train the image data globally (global method), we introduce the local CNN method before the feature extraction by deep learning. The purpose of introducing this new method in deep learning pipeline is to investigate whether the pre-processing can contribute to the performance of the deep learning approach. The term

'new local CNN' refers to the new local technique introduced into the transfer learning pipeline. Previously, studies used global CNN technique that directly processes the whole image frame into feature extraction by transfer learning model. However, in this study, we introduced the 'new local CNN' technique before feature extraction by transfer learning model. To obtained the localize data information, we cut the globalize image into two equal localize images (upper and lower part). For a clear explanation about the methodology, refer Fig. 4.

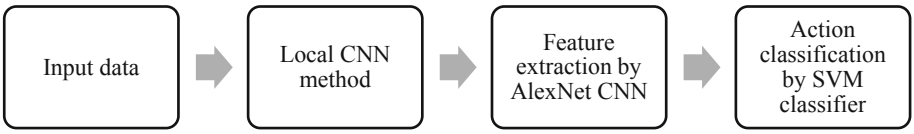

Fig. 3. The block diagram of the methodology

At the beginning of this study, each localize images were trained separately using the pre-trained AlexNet CNN to obtain the features for upper localize and lower localize images. The training took place on Nvidia GeForce GT 740 with computing capability 3.0 using Matlab 2018b. Next, both features from upper and lower localize images that were extracted at fc8 layer were classified using SVM classifier to predict the action class either hit or non-hit action. The performance of the proposed method was analysed and illustrated in form of confusion matrix and accuracy table. The experiment was repeated using normal globalize CNN feature extractor to compare and observe the difference on their performance accuracy.

4 Results and Discussion

Table 2 shows the accuracy table for classification task of both global and local CNN feature extractor trained using AlexNet model and classified using SVM classifier. Interestingly, for our proposed local CNN method, the performance accuracy was found higher compared to the traditional global CNN method. In our view, the result empahizes the validity of our method. Generally, CNN process the whole raw image frame pixels to extract the features. However, with our proposed method, instead of extensively process the whole image frame, the global image frame was divided into two equal local parts. We believe that the local method is more accurate because it carries more informations since each local part containing smaller background area or pixels as compare to processing whole image frame globally. Thus, the local CNN is susceptible to the influence of background pixels.

Figures 5 and 6 show the confusion matrix of global and local CNN feature extractor. From Fig. 5, 5 hit actions were falsely classified as non-hit action and 8 non-hit actions were falsely classified as hit action and it make a total of 13 images were falsely classified (2.2% of misclassification). As in Fig. 6, from a total of 299 hit actions, 6 were falsely classified as non-hit and from 299 non-hit actions, 2 were falsely classified as hit. A total of falsely classified actions are only 8 images which make the accuracy is higher compared to the previous one (1.3% of misclassification).

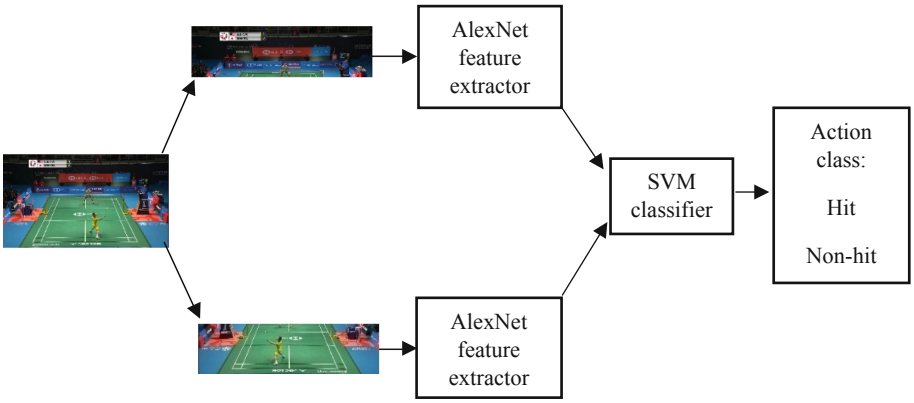

Fig. 4. The illustration of experimental setup

Table 2. Accuracy table

Method for classification task	Accuracy (%)
Globalize CNN feature extractor + SVM	97.8
Localize CNN feature extractor + SVM	98.7

Fig. 5. Confusion matrix of global CNN feature extractor

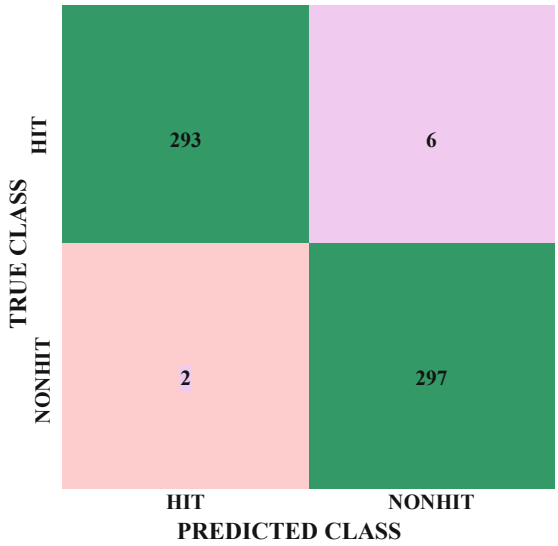

Fig. 6. Confusion matrix of new local CNN feature extractor

5 Conclusion

As a conclusion, we found out that the local CNN extractor introduced in the deep learning pipeline as a part of pre-processing can improve the performance action recognition. Eventhough there are a lot of study have been done by other researchers in action recognition using deep learning approach, a lot of improvement is still needed. Hence, we believe that our proposed method can be used to enhance the performance of action recognition. In future, using our proposed method, we will study more on classifying the specific actions in badminton.

Acknowledgement. The authors would like to express their gratitude to Universiti Teknologi Malaysia (UTM) and the Minister of Education (MOE), Malaysia for supporting this research work under Zamalah UTM and FRGS Research Grant No. R.J130000.7851.5F108.

References

1. Cust, E.E., et al.: Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. J. Sports Sci. **37**(5), 568–600 (2019)
2. Zerrouki, N., et al.: Vision-based human action classification using adaptive boosting algorithm. IEEE Sens. J. **18**(12), 5115–5121 (2018)
3. Tejero-de-Pablos, A., et al.: Human action recognition-based video summarization for RGB-D personal sports video. In: 2016 IEEE International Conference on Multimedia and Expo (ICME) (2016)

4. Han, Y.M., et al.: Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recogn. Lett. **107**, 83–90 (2018)
5. Karpathy, A., et al.: Large-scale video classification with convolutional neural networks, pp. 1725–1732 (2014)
6. Sozykin, K., et al.: Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks. Computer Vision and Pattern Recognition, abs/1709.01421 (2017)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition CoRR, abs/1409.1556 (2014)
8. Qin, J., et al.: Compressive sequential learning for action similarity labeling. IEEE Trans. Image Process. **25**(2), 756–769 (2016)
9. Baek, J., Yun, B.-J.: Posture monitoring system for context awareness. Mob. Comput. **59**, 1589–1599 (2010)
10. San-Segundo, R., et al.: Human activity monitoring based on hidden Markov models using a smartphone. IEEE Instrum. Meas. Mag. **19**(6), 27–31 (2016)
11. Tao, Y., Hu, H.: A novel sensing and data fusion system for 3-D arm motion tracking in telerehabilitation. IEEE Trans. Instrum. Meas. **57**, 1029–1040 (2008)
12. Yu, L., et al.: Nonintrusive appliance load monitoring for smart homes: recent advances and future issues. IEEE Instrum. Meas. Mag. **19**(3), 56–62 (2016)
13. Mlakar, M., et al.: Analyzing tennis game through sensor data with machine learning and multi-objective optimization. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, pp. 153–156. ACM, Maui (2017)
14. Renò, V., et al.: A technology platform for automatic high-level tennis game analysis. Comput. Vis. Image Underst. **159**, 164–175 (2017)
15. Cheng, J., et al.: Designing sensitive wearable capacitive sensors for activity recognition. IEEE Sens. J. **13**(10), 3935–3947 (2013)
16. Gaddam, A., Mukhopadhyay, S.C., Gupta, G.S.: Elder care based on cognitive sensor network. IEEE Sens. J. **11**(3), 574–581 (2011)
17. Kan, Y., Chen, C.: A wearable inertial sensor node for body motion analysis. IEEE Sens. J. **12**(3), 651–657 (2012)
18. Wang, J., et al.: Learning actionlet ensemble for 3D human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 914–927 (2014)
19. Saif, S., Tehseen, S., Kausar, S.: A survey of the techniques for the identification and classification of human actions from visual data. Sensors **18**(11), 3979 (2018)
20. Yoshikawa, F., et al.: Automated service scene detection for badminton game analysis using CHLAC and MRA. World Acad. Sci. Eng. Technol. **4**, 841–844 (2010)
21. Zhao, Y., et al.: Recognizing human actions from low-resolution videos by region-based mixture models. In: 2016 IEEE International Conference on Multimedia and Expo (ICME) (2016)
22. Shishido, H., et al.: A trajectory estimation method for badminton shuttlecock utilizing motion blur. In: Image and Video Technology, pp. 325–336. Springer, Heidelberg (2013)
23. Ma, Z., Sun, Z.: Time-varying LSTM networks for action recognition (2018)
24. Yang, J.B., et al.: Deep convolutional neural networks on multichannel time series for human activity recognition. In: IJCAI, Buenos Aires, Argentina (2015)
25. Alex, K., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks, pp. 1097–1105 (2012)
26. Shih, H.: A survey on content-aware video analysis for sports. IEEE Trans. Circuits Syst. Video Technol. **28**(5), 1212–1231 (2017)

27. Tora, M.R., Chen, J., Little, J.J.: Classification of puck possession events in ice hockey. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
28. Baccouche, M., et al.: Action classification in soccer videos with long short-term memory recurrent neural networks. In: Artificial Neural Networks – ICANN 2010: 20th International Conference. Springer, Heidelberg (2010)
29. Ullah, A., et al.: Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access **6**, 1155–1166 (2018)