# The new Convolutional Neural Network (CNN) local feature extractor for automated badminton action recognition on vision based data

# The new Convolutional Neural Network (CNN) local feature extractor for automated badminton action recognition on vision based data

**N A Rahmad [1] and M A As'ari [*2]**

[1] School of Biomedical Engineering and Health Sciences, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.
[2] Sport Innovation and Technology Center (SITC), Institute of Human Centered Engineering (IHCE), Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

amir-asari@biomedical.utm.my

**Abstract**. Automated action recognition is useful for improving the performance of the athletes through notational analysis. The notational analysis is usually used by the coach or notational analyst to study the movement patterns, strategy and tactics. Therefore, action recognition is the main key before further analysis can be done. This paper focused on developing an automated badminton action recognition using vision based dataset. 1496 badminton match image frames of 5 actions were studied – smash, clear, drop, net shot and lift. At first, the dataset was classified into 0.8:0.2 for training and testing the classification task by machine learning. Secondly, features of the training dataset were extracted using the Alexnet Convolutional Neural Network (CNN) model. In extracting the features, we introduced the new local feature extractor technique that extracts features at the fc8 layer. After collecting all the features at the fc8 layer, features were being classified by using machine learning classifier which is linear Support Vector Machine (SVM). The experiment was repeated using a normal global feature extractor technique. Lastly, both of the new local and global feature extractor techniques were repeated using GoogleNet CNN model to compare the performance between AlexNet and GoogleNet model. The results show that the new local feature extractor using AlexNet CNN model has the best performance accuracy which is 82.0%.

**Keywords:** action recognition, convolutional neural network, deep learning, sport performance analysis, vision based.

## 1. Introduction

Human action recognition (HAR) is not new in computer vision area as it has been studied by many researchers before. It is beneficial in many applications because the interest in HAR system has reported to be increased [1] especially on vision based data. The example of applications that used action recognition system are video surveillance [2], human monitoring for healthcare related area [3], human-computer interaction [4] and sport performance analysis [5]. The main focus of this study is to utilize the automated action recognition system for sport performance analysis application.

Action recognition is the main key for performance analysis because actions need to be recognized first before further analysis such as tactical assessment, movement analysis and opposition performance can be done. However, the current system is mainly depending on the notational analyst

to manually interpret live match and create video playlists for performance analysis tasks which is time consuming and not practical. Therefore, with the introduction of an automated action recognition system in sport field, it can overcomes the stated problems.

The video based action recognition is a current trend among researchers due to an advance development of technology nowadays that make broadcast video of sports match can be obtained online easily. We are analyzing the broadcasted video of badminton match to make the technology in badminton to be in the same level as technology in other sports such as football and basketball.

Since the introduction of deep learning in computer vision area, more researches have focused on this approach. It is a subtype of machine learning, but promotes favorable results compared to the machine learning method. Unlike machine learning, the pipeline does not involve manual feature learning and extraction because relevant features are automatically extracted from the input image frames. There are 3 common ways to use deep learning for classification or action recognition task. First is by training from scratch which required thousands of input data and powerful hardware graphical processing unit (GPU). Secondly, by transfer learning which mean fine-tuning the pre-trained model such as AlexNet [6], GoogleNet, VggNet [7] and so on with own dataset. This approach is very useful to train limited input data and also can reduce the training time. Lastly, deep learning is used for more specialized approach which is designing the deep learning model as feature extractor. All the features will be learned and extracted automatically by the network. Therefore, we can pull these features out from network at any desired layer which is then features will be the input for other machine learning.

The objective of this paper is we want to propose a new local feature extractor based on CNN technique using pre-trained deep learning CNN approach which is then the produced features will be classified using the Support Vector Machine (SVM). Two pre-trained models used in this study to extract new local features are AlexNet and GoogleNet model. The performance accuracy of each technique on each model were visualized in the confusion matrix for a better understanding. The next section presents the works related to our study. Then, materials and methods were explained in Section 3. Results, discussion, conclusion and future works also reported accordingly. Overall, this new local feature extractor technique introduced into the deep learning pipeline able to help in improving the performance accuracy of recognition task.

## 2. Related work

A task of recognizing action is crucial in the sports field and has been extensively studied by many researchers. According to [5], 80% of the current works are mainly focusing on football, basketball [8] and tennis [9-10] while another 20% are among other sports such as hockey, badminton and swimming [11].

Previously, researchers used the handcrafted feature extraction and classification using the machine learning technique to fulfill the recognition task. For video based modality, the study in [12] stated that features can be in form text, audio or visual. For instance, work in [13] classifies sports genre (basketball, soccer and tennis) by using pseudo-2D-Hidden Markov Model (HMM) classifier from the low-level visual and audio features. A study by [14] also proposed visual features that were manually extracted 10 computable spatio-temporal features and classified using hierarchical SVM to deal with multiclass problem. Not only that, research done by [15] presents a classification task of 14 sports genre on huge-scale and low resolution sports video obtained from online video sources. The study formulated handcrafted feature extraction method which is bag of visual-words using speeded-up robust features (SURF) and classified using SVM.

The development of technology and the availability of the dataset has led to the introduction of deep learning which required enormous dataset. The current trend among researchers is the study of learning the features from the trained deep neural network for action recognition. For instance, the work in [16] determined the features on standard video actions by proposing two streams ConvNet architecture of spatial and temporal networks. It showed that by using the pre-trained network, it could improves the performance accuracy compared to train the network from scratch. Work by [17] also

utilized deep CNN based method to extract visual features by fine-tuning the deep CNN and fed features into Recurrent Neural Network (RNN) to draw the visual information of the video caption for action recognition purpose. The similar study done by [18] on an ice hockey dataset in which extracting features from the whole frame using pre-trained AlexNet CNN and then used Long-Short Term Memory (LSTM) model to train an event classification from the extracted features.

## 3. Materials and methods

This study was conducted on our own constructed video based dataset. It comprises of 1496 image frames that were extracted using VirtualDub software from 5 badminton broadcasted videos obtained from BadmintonWorld.tv channel on Youtube database. All the 1080p High Definition (HD) resolution image frames are from the single player rally scene with the top behind the backline view. Table 1 shows the details of our dataset used in this study.

**Table 1.** The distribution of dataset.

| Action | No. of total sample | No. of training sample | No. of testing sample |
|---|---|---|---|
| Clear | 200 | 160 | 40 |
| Drop | 100 | 80 | 20 |
| Lift | 398 | 318 | 80 |
| Net shot | 270 | 216 | 54 |
| Smash | 528 | 422 | 106 |
| **Total** | **1496** | **1196** | **300** |

Figure 1 describes how the badminton actions were recognized and classified in this study using computer vision algorithm. These experiments that consist of training and testing were fully conducted on Matlab 2018b software using the training options shown in Table 2.

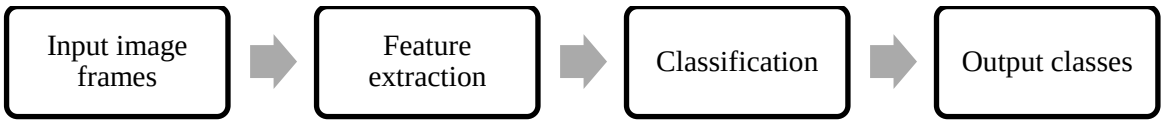


**Figure 1.** The block diagram of the experimental flow.

**Table 2.** Training options.

| Training options | |
|---|---|
| Training optimizer | sgdm |
| Mini-batch size | 5 |
| Maximum epochs | 10 |
| Execution environment | gpu |
| Initial learning rate | 0.0001 |

As for the feature extraction task, pre-trained CNN models were used to automatically learn and extract relevant visual features of each image frame at fc8 layer of AlexNet model and fc of GoogleNet model. The features that were automatically learned inside the deep networks consist of low-level features such as edges and colors to high-level representation of the image. Generally, deep CNN learns the whole image frames that carry global features or information. However, as for new local feature extractor technique, it only learns the local features from each smaller part of the whole frames. We introduced pre-processing before deep CNN is doing the feature extraction using this new local technique as shown in Figure 2. Basically, this pre-processing divided the whole image frames

into two parts – upper and lower in which each part carries their own local features or information. Figure 3 shows the sample of global image frame while Figure 4 shows the sample of local image frames. After feature extraction, all features were being classified using the machine learning classifier which is linear SVM before attaining the five output classes.

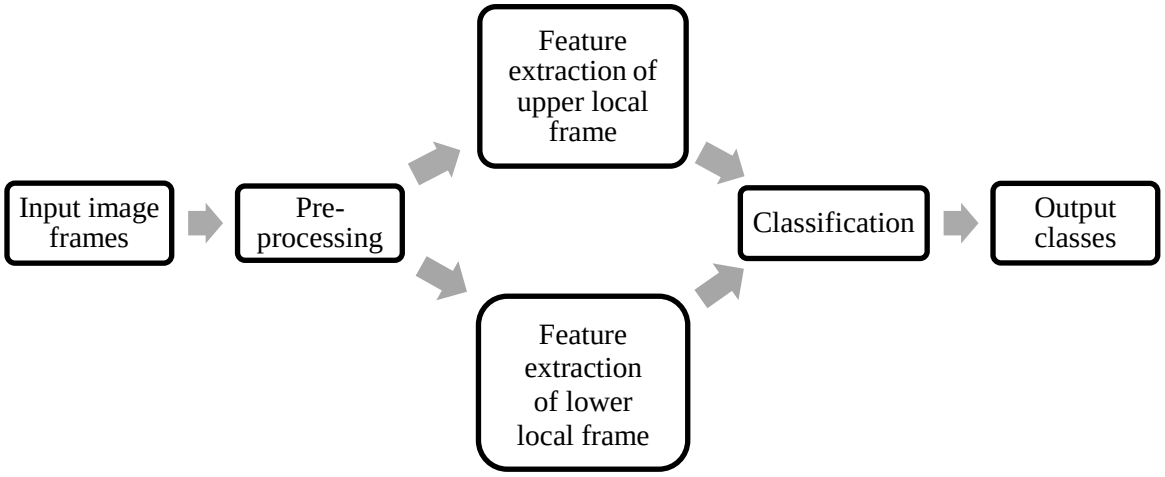


**Figure 2.** The block diagram of the new local feature extractor experimental flow.


**Figure 3.** Global input image frame.

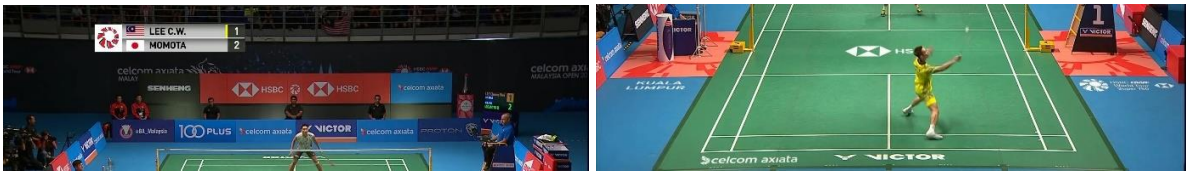


**Figure 4.** Local input image frames.

## 4. Results

Table 3 shows the accuracy table in recognizing badminton actions by the linear SVM classifier using local and global feature extractor method for both AlexNet and GoogleNet models. The best performance accuracy for each technique is shown in bold. Furthermore, Figure 5 until Figure 8 is the confusion matrix of each new local and global feature extractor technique using AlexNet and GoogleNet models. Row represents predicted or output class while column represents an actual target class. All the correct predicted classes are in the diagonal. From the figures, 1- clear, 2- drop, 3- lift, 4- net shot and 5-smash.

**Table 3.** Accuracy table.

| Techniques | Accuracy (%) | |
|---|---|---|
| | **AlexNet model** | **GoogleNet model** |
| Local | **82.0** | 68.0 |
| Global | 66.7 | **85.7** |


**Figure 5.** Confusion matrix of new local feature extractor technique (AlexNet model).

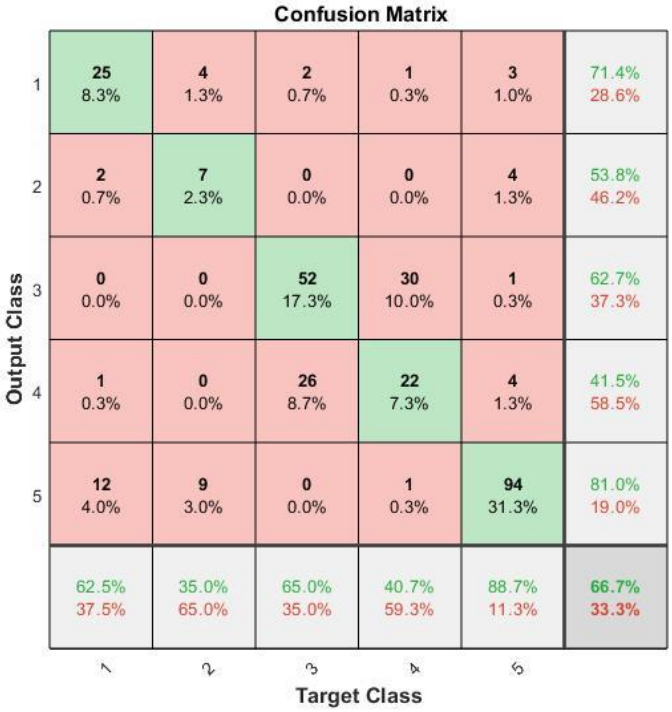


**Figure 6.** Confusion matrix of global feature extractor technique (AlexNet model).


**Figure 7.** Confusion matrix of new local feature extractor technique (GoogleNet model).

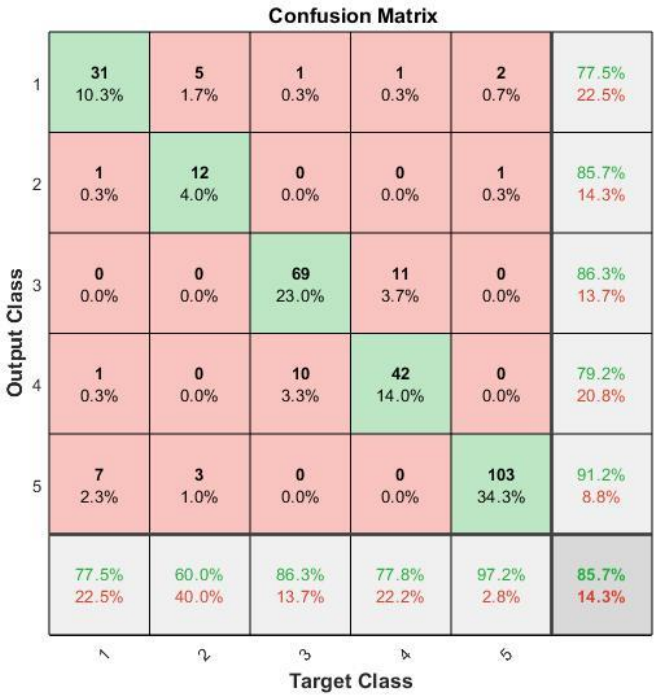


**Figure 8.** Confusion matrix of global feature extractor technique (GoogleNet model).

## 5. Discussion

As we can see from the results, it is clearly shown that the performance accuracy varies depending on feature extraction techniques. Interestingly, new local feature extractor technique performed much way better than global technique for AlexNet model meanwhile GoogleNet model shows otherwise. The highest accuracy of the new local feature extractor technique is from AlexNet model which is 82.0% while the highest accuracy of global feature extractor technique is from GoogleNet model which is 85.7%.

Our experiments proved that the proposed new local feature extractor is very useful towards improving the performance of classification tasks. This is because local image frames carry more information than the whole image frames. It has two types of information – 1) The action done by the player and 2) The action made by opponent responding to the action done by the player. However, we found much lower value for performance accuracy of new local feature extractor technique than to global feature extractor technique for GoogleNet model.

From confusion matrix, we can interpret that the smash action contributes the most to the performance accuracy of each cases and drop action has the least contribution. We might say that the unbalance data distribution influence the performance accuracy. Moreover, as for a deeper GoogleNet model, the model might be underfit in which it cannot generalize new data and produces a lower performance accuracy of the new feature extractor technique because there is not enough data to train the deeper GoogleNet model.

## 6. Conclusion and future works

In order to automatically recognize badminton action for performance analysis purpose, several feature extraction techniques have been explored to extract the features from the image frames and linear SVM classifier has been used to classify the features. From the study, we found out that this new local feature extractor technique is the best technique that can be used to improve the performance accuracy of classification tasks. In a conclusion, the proposed new local feature extractor can be considered as

our contribution and the novelty of this study includes the introduction of the pre-processing task before the feature extraction task in the deep learning pipeline. In future, more experiments need to be done such as conducting an experiment on more dataset or conducting experiments on features extracted at different fully connected layer to properly investigate the root cause behind the obtained results.

**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] Aggarwal J K and Ryoo M S Human activity analysis: a review *ACM Computing Surveys (CSUR)* 2011 43
[2] Hu W, Tan T, Wang L and Maybank S A survey on visual surveillance of object motion and behaviors *IEEE Transactions on Systems, Man, and Cybernetics,Part C: Applications and Reviews* 2004 34 334–52.
[3] Mukhopadhyay S C Wearable sensors for human activity monitoring: a review *IEEE Sensors Journal* 2015 15
[4] Jaimes A and Sebe N Multimodal human-computer interaction: a survey *Computer Vision and Image Understanding* 2007 108 116-34.
[5] Shih H C A survey of content-aware video analysis for sports *IEEE Transactions on Circuits and Systems for Video Technology* 2017 28 1212-31.
[6] Alex K, Sutskever I and Hinton G E ImageNet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* 2012 1097-105.
[7] Simonyan K and Zisserman A Very deep convolutional networks for large-scale image recognition *preprint* arXiv:14091556 2015
[8] Bertasius G, Park H S, Yu S X and Shi J Am i a baller? basketball performance assessment from first-person videos *preprin*t arXiv:161105365 2017
[9] Renòa V, Moscaa N, Nittia M, D'Orazioa T, Guaragnellab C, Campagnolic D, Pratid A and Stellaa E A technology platform for automatic high-level tennis game analysis *Computer Vision and Image Understanding* 2017 159 164-75.
[10] Sukhwani M and Jawahar C V 2016 *Frame level annotations for tennis videos* (Cancun, Mexico) 841-46.
[11] Sha L, Lucey P, Morgan S and Pease D 2013 *Swimmer localization from a moving camera* (Hobart, Australia) 1-8.
[12] Brezeale D and Cook D J Automatic video classification: a survey of the literature *IEEE Trans Syst Man Cybernetics—PART C: Applicant Reviews* 2008 38 416–30.
[13] Wang J, Xu C and Ching E 2008 *Automatic sports video genre classification using pseudo–2d–HMM* (Hong Kong, China) 778-81.
[14] Yuan X, Lai W, Mei T, Hua X S, Wu X Q and Li S 2006 *Automatic video genre categorization using hierarchical SVM* (Georgia, USA) 2905-08.
[15] Li L, Zhang N, Duan L Y, Huang Q, Du J and Guan L 2009 *Automatic sports genre categorization and view-type classification over large-scale dataset* (Columbia, Canada) 653–56.
[16] Simonyan K and Zisserman A 2014 *Two-stream convolutional networks for action recognition in videos* (Montreal,Canada) 1-11.
[17] Li G, Ma S and Han Y 2015 *Summarization-based video caption via deep neural networks* (Brisbane, Australia) 1191–94.
[18] Tora M R, Chen J and Little J J 2017 21-26 July 2017 *Classification of puck possession events in ice hockey* (Hawai, USA) 147-54.