Sports Action Classification using Deep Learning Model

Shayla Sharmin
Department of Computer and Information Sciences
University of Delaware, Newark, DE-19713
shayla@udel.edu

Prerana Khatiwada
Department of Computer and Information Sciences
University of Delaware, Newark, DE-19713
preranak@udel.edu

Abstract

Recognizing human activity entails correctly anticipating a human's action, which necessitates a set of data points. It is important in different area in computer science and in real life, for example, sports classification. Sports are a popular kind of entertainment and a way of living for sportsmen in the modern world. Post-game analysis, automatic scoreboard updates, and even automated commentary can all benefit from sports classification Using CNN (convolutional neural networks) to learn features or visual cues and RNN (recurrent neural networks) to identify the temporal relationship between frames, a proposed model has been trained using the KTH dataset that can distinguish six classes of action i.e. boxing, hand-clapping, hand waiving, jogging, running, and walking. The average accuracy for the training data set is 98.672% and 54.31% for test dataset.

1 Introduction

A video classifier should be able to label video frames reliably, characterize their properties, and annotate them. In video classification, human activity recognition is an intriguing topic. Human activity recognition implies correctly predicting a human's action, which requires a set of data points. In this work, Russo et al., [1] work has been implemented. Russo et al., [1] proposed a sports video classification that can distinguish football, cricket, tennis, basketball, and ice hockey into five distinct sports based solely on visual information. A total of 300 video frames drawn from 50 different YouTube videos have been trained in their proposed model and achieved 99.58% accuracy on the training dataset and 96.66% accuracy on the test dataset. In this work, We select sports classification, because it can assist viewers in viewing relevant items while watching games, as well as coaches and players. Aside from that, there are some real-world sports applications, such as context-based advertisement, highlight extraction, match summary, and automatic commentary generation.

Due to its commercial benefits and a wide range of viewers, sports-based research is prevalent nowadays. In recent years, several methods aim to classify different sports actions from videos. Russo et al. [1], [2] proposed a model to classify sports using video. In their works, they used their own dataset of various sports. . Sen et al., [3] proposed a hybrid deep-neural-network architecture for classifying 10 different cricket batting shots from offline videos. Sen et al., also categorized six different soccer actions in 2021 [4]. The authors of [5] employed thermal imaging to create heatmaps, projecting them onto a low-dimensional space with two separate approaches: PCA and Fischer's Linear Discriminant. For five classes, their experiment and the results was promising. In

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

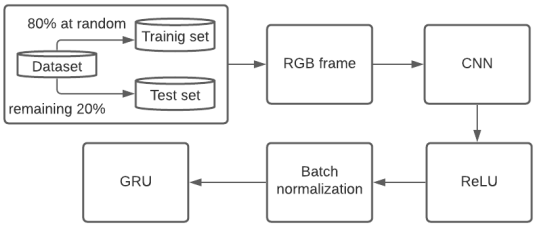


Figure 1: Proposed methodology

[6], both audio and video characteristics were combined for sports classification,and the results were satisfactory. Bhalla et al., [7] develops a cricket game summarization system that uses video shot detection, sound detection, and scoreboard recognition. Rafiq et al., [8] presented a method based on transfer learning for classifying sports video sequences in order to provide video summarization. The suggested Alexnet CNN-based technique achieves a 99.26 percent accuracy over a smaller dataset including solely cricket scenes.

Rangasamy et al.,[9] proposed a VGG16-based deep learning model for classifying four field hockey actions, including free hits, goals, penalty corners, and long corners. Due to the lack of a publicly available dataset, a new dataset is created. The highest accuracy of 98 percent is achieved after 300 epochs of training. Sarma et al.,[10] suggested using a deep learning approach to determine five traditional Bangladeshi sports, including Boli Khela, Kabaddi, Lathi Khela, Kho Kho, and Nouka Baich. Open source YouTube videos are used to generate a new video dataset. The hybrid VGG19–LSTM model achieves 99 percent accuracy. Steel et al.,[11] introduced a method for distinguishing 9 different badminton actions by using a convolutional neural network and accelerometer data. The weighted accuracy of the sensor placement in the wrist and upper arm was 93 percent and 96 percent, respectively, which was somewhat less than the average accuracy of 98 percent. Junjun et al., [12] proposed a method for categorizing between three different varieties of basketball actions: shot, pass, and catch.

So far, we have learned about the paper's motivation, application, and some related papers. We will go over the model's problem formulation, proposed approach, and conclusion after looking at the numerical results in the upcoming sections.

2 Problem formulation or model

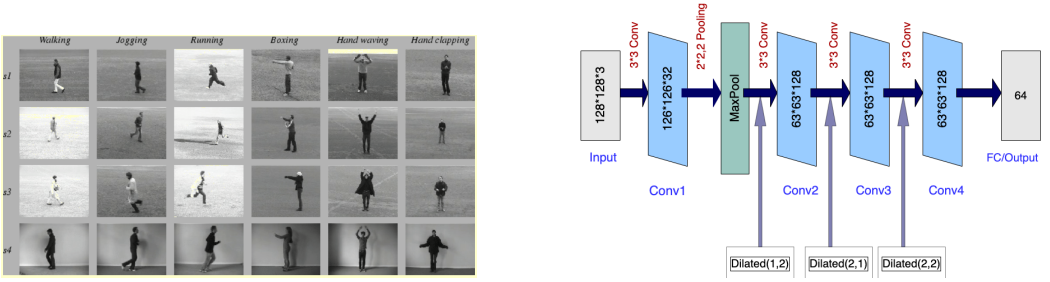
The aim of this research is to implement paper of Russo et al., [1] to investigate the effectiveness of their proposed model using KTH data set. KTH dataset has six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) [13] where Russo et al., classified their own dataset which has five sports football, cricket, tennis, basketball, and ice hockey [1]

3 Proposed Approach

The proposed approach can be divided into two groups, i.e. Convolution Neural Network and Recurrent Neural Network. This paper followed the methodology proposed by Russo et al.,[1] where they performed several experiments and kept the best one. Figure 1 shows the process that has been followed in this paper.

3.1 Dataset

Russo et al., used their own dataset to run their model that contains 300 video sequence from 50 different videos [1]. To test the model they run the model several times and go and From the dataset 80% are selected as training data randomly and remaining 20% are consider as the test set. 2a shows a sample of the data set that has been used in this paper.



(a) KTH Database [13]

(b) Architecture of the CNN Model [1]

Figure 2: (a) Dataset (b) CNN model

3.2 RGB frames

These video sequences are converted in RGB color frames and used input of the convolution network. Russo et al., [1] got 99.58% for training dataset using 8 frames in a sequence and 96.66% for test dataset using 6 and 13 frames in a sequence. In this work, 13 frames in a sequence is used for the experimental result analysis.

3.3 Convolutional Neural Network

Convolutional neural network (CNN) is a version of multi layer perceptrons that includes convolutional layer, non-linearity layer, pooling layer and fully-connected layer[14]. Fully connected layer maps the features that are extracted in convolution layer and pooling layer []. This is designed to automate the spatial feature extraction process from images. In the convolution step, the input image is convolved with some filter and rolled over the whole image. The movement of filters across the whole image can be controlled by specifying a value for stride. To overcome the downside of shrinking outputs and losing information, padding value can be set. The convolution layer has four parts which has been shared in figure 2b. Each layer share their weights. The ***first layer*** consists of a 32 feature map that has 3 x 3 receptive field. And after that there is a **maxpool** with stride 2. Max pooling is used to aid over-fitting by providing an abstracted representation of the data. It also lowers the computational cost by minimizing the number of parameters that must be learned and gives basic translation invariance to the internal representation [15]. Maximum pooling extracts patches from the input feature maps, outputs the maximum value in each patch, and discards all the other values. The next three layers are three different dilated convolutional layer. The description is given in table 1.

Table 1: Description of Convolution Part

Layer	1	2	3	4
Filter Size	3 x 3	3 x 3	3 x 3	3 x 3
Dilation (width, height)	-	(1,2)	(2,1)	(2,2)
Feature Map	32	128	128	128
Non Linearity	ReLU	ReLU	ReLU	ReLU
Resizing	Max pooling			

Dilated Convolution is a variation of convolution [16] where it uses a defined gap. For a convolution operation to be termed as dilated convolution, the factor of the gap must be greater than or equal to 2. When the dilation factor is 1, the operation is referred to as a traditional convolution operation. For example a video of running or jogging contain objects other than only participant's action. While walking surrounding can be an important aspect of defining the cricket shot. Thus, to have an larger overview of the scene content, dilated convolution is used here. After the dilated convolution layer, the next layer is a **fully connected layer** with 64 neurons. This network receives the output of the CNN model and sends to the next step that is recurrent neural network. The convolutional layers' output represents the data's high-level characteristics. While the output layer could be flattened and connected to the output layer, adding a fully-connected layer provides a (usually) inexpensive

means of learning non-linear combinations of these features [17]. The fully-connected layer learns a (potentially non-linear) function in that space, with the convolutional layers giving a meaningful, low-dimensional, and fairly invariant feature space. The number of output nodes in the final fully connected layer is usually equal to the number of classes. A nonlinear function, such as ReLU, follows each fully connected layer [17].

3.4 Recurrent Neural Network

A recurrent neural network (RNN) is a kind of artificial neural network in which links between nodes form a graph that is guided along a time sequence. Unlike conventional neural networks where the inputs and outputs are independent of each other, in RNN, the previous step's output is fed as an input to the current step. The foremost important feature of RNN is its hidden state, which recalls some sequence information. RNNs use their internal state (memory) motivated by the feed forward neural networks to process varying input length sequences. This concept can be applied to various time-dependent applications such as speech recognition, video sequencing, machine translation, music composition, etc.

The activation function, which is placed at the output of the CNN layers, is a rectified linear unit (ReLU). Activation function helps to decide if the neuron would activate or not. Depending upon the function, it is used to map the resulting values in between 0 to 1 or -1 to 1 etc. After each activation, batch normalization was performed with RMSProp as the Optimizer. During training, batch normalization makes the network more stable [18]. This may necessitate the usage of considerably higher than normal learning rates, which could speed up the learning process even further. The learning rate is 0.0001, the loss function is category cross entropy, and there are 60 epochs in this example. In addition, gated recurrent units for the recurrent component (GRU) were used in this study introduced by Cho et al., [19]. GRU is a simpler form of LSTM developed by researchers. GRU has demonstrated its unrivaled performance in long-term feature reliance using fewer gates than LSTM. Table 2 shows the information of the model that have been followed from the selected paper [1].

Table 2: Model information [1]

Activation function	ReLU
Normalization method	Batch normaliztion
Learning rate	0.0001
Optimizer	RMSProp
Loss function	Categorical cross entropy
Frame	13 per video
Epoch	60

4 Numerical Experiment

The model proposed by the Russo et al., [1] has been tested using KTH database. Google Colab is used to run the model. This model's dataset and result analysis are explained in the subsections that follow. This section explains how to evaluate the given model using various strategies. To compare our suggested architectures, we employed precision, recall, and F1–score accuracy measurements. In addition, the accuracy and loss curves obtained from the model are shown here.

4.1 KTH data set for sports

For the experiment we use video database from KTH that contains six different types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed multiple times by 25 subjects in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3, and indoors s4, as shown below [13]. From the dataset 80% has been taken as train dataset while the rest of the 20% considered as test dataset

Table 3: Train and Test set of KTH dataset

Class	Boxing	Handclapping	Handwaving	Jogging	Running	Walking	Total
Test	20	19	20	20	20	20	119
Train	80	80	80	80	80	80	480

4.2 Result Analysis

We run the model five times in this experiment and get the following result displayed in table 4. The train and the test set were in random order. Each iteration's split was done at random, with 80% of the training set and 20% of the test set. The average training and test accuracy is 98.672% and 54.31%, respectively. From the result we can say that for KTH dataset the model is overfitting. Because we have high training accuracy in comparison to test accuracy. We have included the classification report and confusion matrix for iteration 4 in the following section because we had good test accuracy here.

Table 4: Average Accuracy of Training and Test Result

Iteration	1	2	3	4	5	**avg**
Train accuracy%	99.19	100	100	94.17	100	98.672
Test accuracy%	53.13	56.35	55.05	57.14	49.88	54.31
epoch	52	47	43	45	49	-

Confusion matrix

The confusion matrix for the iteration 4 the model is depicted in Figure 3. Handclapping has the highest number of correctly classified data (16 out of 19), whereas hand waving has the lowest number of correctly classified data (5 out of 19). Thirteen of the twenty students in the hand waving class were misclassified as hand clapping. In the confusion matrix, the right classifications for boxing, jogging, running, and walking are 13, 9, 10, and 7, respectively.

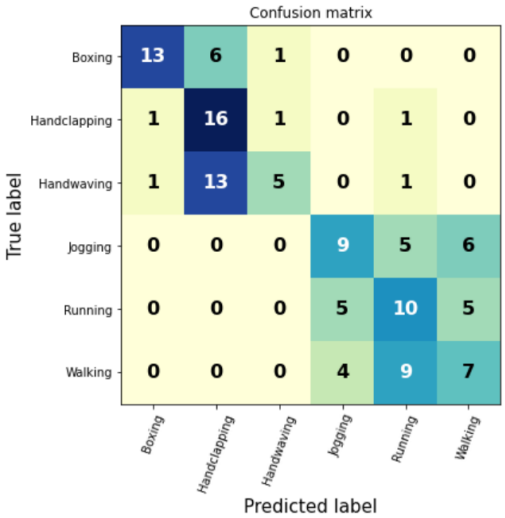


Figure 3: Confusion Matrix

Classification report

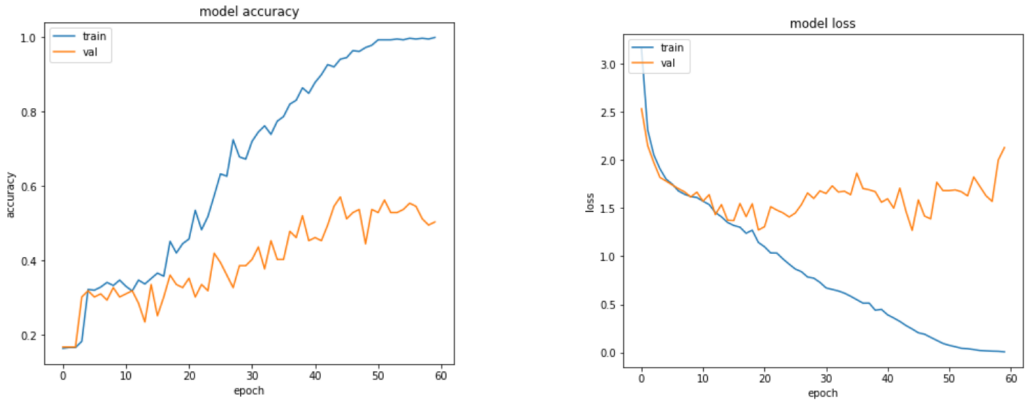
The classification report for the proposed model [1] for KTH data set is summarized in table 5. The model calculates a low precision value of 38% for the class "Running" among the six actions. Because 13 handwaving was mistaken for handclapping, it had the lowest recall value of 25%. The greatest f1 score value is 74 percent for "Boxing class," while the lowest is 37 percent for handwaiving and walking, with 15 and 13 miss classifications, respectively. The average f1 score is 50%.

Table 5: Classification Report

Class	precision	recall	f1-score	support
Boxing	0.87	0.65	0.74	20
Handclapping	0.46	0.84	0.59	19
Handwaving	0.71	0.25	0.37	20
Jogging	0.50	0.45	0.47	20
Running	0.38	0.50	0.43	20
Walking	0.39	0.35	0.37	20
f1 score accuracy			0.50	119
macro avg	0.55	0.51	0.50	119
weighted avg	0.55	0.50	0.50	119

Accuracy Curve and Loss Curve

Figure 4(a) depicts the accuracy curve for the proposed model [1], whereas Figure 4(b) depicts the loss curve during the training process. Both the figures are shown for 60 epochs. The highest of 57.14% test accuracy is recorded on 45^{th} epoch with training accuracy 94.17 %. The highest training accuracy is 100% at 60^{th}. 60^{th} epoch but the test accuracy decrease to 50.02% from 57.14%. In 45^{th} epoch the loss value of training is 0.24 and loss value of test is 1.27.



(a) Accuracy curve

(b) Loss curve

Figure 4: Accuracy and Loss curve

Conclusion

Sports video categorization analysis has commercial implications for broadcasters, as well as the necessity of generating sensor-based unbiased commentary and helping coaches train and improve their players' tactics. In this work, we implement Russo et al., [1] using KTH dataset. Not just human acts, but also the surrounding environment, play a vital role in classifying sports. To categorize these types of behaviors, high-level features are required. Large input sizes or deeper neural networks can be used to achieve high-level features, but this requires high-configuration hardware. KTH dataset is a simple dataset and shows 57.14% accuracy for test set and 94.17% accuracy for training set in 45^{th} epoch while compiling the model in iteration 4. The average training accuracy is 98.672% and average test accuracy is 54.31%. Russo et al., got 96.66% test accuracy for their own dataset. Although the author of this paper got high test and train accuracy for their own dataset, for kth dataset it is overfit. As a part of our future work, more sports categories with their various activities could be added along with solving the overfitting problem.

References

[1] Russo Mohammad Ashraf Uddin, Alexander Filonenko, and Kang-Hyun Jo. Sports classification in sequential frames using cnn and rnn. *2018 International Conference on Information and*

Communication Technology Robotics (ICT-ROBOT), pages 1–3, 2018.

[2] Mohammad Ashraf Russo, Laksono Kurnianggoro, and Kang-Hyun Jo. Classification of sports videos with combination of deep learning models and transfer learning. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–5, 2019.

[3] Anik Sen, Deb Kaushik, Pranab Dhar, and Takeshi Koshiba. Cricshotclassify: An approach to classifying batting shots from cricket videos using a convolutional neural network and gated recurrent unit. *Sensors*, 21:2846, 04 2021.

[4] Anik Sen and Deb Kaushik. Categorization of actions in soccer videos using a combination of transfer learning and gated recurrent unit. *ICT Express*, 03 2021.

[5] R. Gade and T. Moeslund. Sports type classification using signature heatmaps. *IEEE*, 2013.

[6] MadsChristensen Rikke Gade, Abou-Zleikha and Thomas Moeslund. Audio-visual classification of sports types. *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.

[7] Aman Bhalla, Arpit Ahuja, Pradeep Pant, and Ankush Mittal. A multimodal approach for automatic cricket video summarization. In *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 146–150. IEEE, 2019.

[8] Muhammad Rafiq, Ghazala Rafiq, Rockson Agyeman, Gyu Sang Choi, and Seong-Il Jin. Scene classification for sports video summarization using transfer learning. *Sensors*, 20(6):1702, 2020.

[9] Keerthana Rangasamy, Muhammad Amir As'ari, Nur Azmina Rahmad, and Nurul Fathiah Ghazali. Hockey activity recognition using pre-trained deep learning model. *ICT Express*, 6(3):170–174, 2020.

[10] Moumita Sen Sarma, Kaushik Deb, Pranab Kumar Dhar, and Takeshi Koshiba. Traditional bangladeshi sports video classification using deep learning method. *Applied Sciences*, 11(5):2149, 2021.

[11] Tim Steels, Ben Van Herbruggen, Jaron Fontaine, Toon De Pessemier, David Plets, and Eli De Poorter. Badminton activity recognition using accelerometer data. *Sensors*, 20(17):4685, 2020.

[12] Gun Junjun. Basketball action recognition based on fpga and particle image. *Microprocessors and Microsystems*, 80:103334, 2021.

[13] Recognition of human actions. `https://www.csc.kth.se/cvap/actions/`.

[14] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, 2017.

[15] Yann Lecun, Patrick Haffner, and Y. Bengio. Object recognition with gradient-based learning. 08 2000.

[16] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[17] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629, 2018.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[19] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

5 Appendix

This work follow the method used in Russo et al. (2005) [1].