ConvLSTM for Table Tennis Stroke Classification

Jansi Rani Sella Veluswami, Ananth Narayanan P, Bhuvan S, Shobith Kumar R

Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India

Abstract

Our study concentrates on sports video analytics, particularly stroke classification. We utilize a model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) trained on the MediaEval Fine-Grained Action Classification of the Table Tennis Strokes dataset. With an accuracy of 81.4%, our model effectively classifies table tennis moves, providing insights for post-match commentary and playstyle analysis. This effectiveness is demonstrated in the context of the MediaEval 2023 benchmark.

1 INTRODUCTION

The field of action recognition involves associating a predefined set of actions with video content to meet the increasing demand for automated action analysis in videos. This paper presents a method that specifically targets the classification of strokes within a dataset of various table tennis strokes performed in match and practice settings. The action recognition process involves localizing objects, identifying them, and then classifying the detected actions. The ability to detect and classify actions is crucial for making strategic decisions, particularly in the context of athletic performance analysis.

The Overview paper [1] describes the dataset TTStroke-21 used in this study which includes 21 different classes of strokes, where two annotated sets are provided: a training and a validation set. Utilizing machine learning in this domain has the potential to enhance athletic performance through computer-aided analysis of moves. In this study, we developed a model implemented using TensorFlow, using a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. Our approach aims to contribute to the improvement of athletic performance by automating the analysis of various strokes. We discuss the results obtained using our model on the given dataset, highlighting the significance of effective action recognition in sports analytics.

2 RELATED WORK

The provided baseline methodology [2] proposes two types of 3D-CNN architectures to solve the subtask. Both the methods are 3D-CNN architectures using Spatio-temporal convolutions and attention mechanisms. The predominant strategies have centered around the utilization of CNN and LSTM-based methodologies. For example, in the paper by Kaustubh Milind Kulkarni *et al.* [3], an LSTM model, a TCN model, and a combined TCN + LSTM model were presented. They used Pose Estimation and a Savitzky-Golay filter for feature extraction.
Kadir Aktas *et al.* [4], present another approach where RGB images were used as the input data without any prior feature extraction. They used an LSTM model to achieve about 79.8% accuracy in validation data. We were inspired by the idea of using the RGB images directly without any feature extraction and executed the same in our work.

MediaEval'23: Multimedia Evaluation Workshop, February 2–3, 2024, Amsterdam, The Netherlands and Online
m.a.larson@tudelft.nl (M. Larson); gareth.jones@computing.dcu.ie (G. Jones); bionescu@imag.pub.ro (B. Ionescu)

CEUR Workshop Proceedings (CEUR-WS.org)

3 APPROACH

A Convolutional Neural Network (CNN or ConvNet) is a type of deep neural network specifically designed for processing image data. This network excels in analyzing images and making predictions based on them. It utilizes kernels, known as filters, to examine the image and generate feature maps, which represent the presence of specific features at various locations within the image. Initially, the network produces a limited number of feature maps, which are augmented and refined through subsequent layers using pooling operations, while retaining critical information without loss.

On the other hand, a Long Short-Term Memory (LSTM) network is specifically designed to handle sequential data, taking into account all previous inputs to generate an output. LSTMs are a type of Recurrent Neural Network (RNN) that addresses the vanishing gradient problem, a limitation of traditional RNNs in handling long-term dependencies in input sequences. This enables LSTM cells to maintain context for extended periods, making them better suited for tasks such as time series prediction, speech recognition, language translation, and music composition.

In the context of action recognition, we will employ a CNN + LSTM network to leverage the spatial-temporal aspects of videos. This combination will enable the network to effectively analyze and recognize actions within video sequences.

3.1 Data Preprocessing

The data preparation process involves class identification and two pivotal functions: one for frame extraction, ensuring resizing and normalization, and another for dataset construction, incorporating features, labels, and video paths. Notably, the dataset creation rigorously filters videos to align with the specified sequence length. The execution of the dataset creation function on a specified directory results in the generation of dataset objects, including features, labels, and video file paths.

These components include features, representing extracted frames from videos, and labels, serving as identifiers for subsequent machine learning model training. The third component consists of paths associated with videos in the dataset, functioning as references to the physical location of each video.

3.2 Proposed Model

The process of creating a dataset for TensorFlow is seamless and incorporates both Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. This choice is informed by the well-established effectiveness of these architectures in video content analysis tasks.

In the construction phase of our model, the Keras ConvLSTM2D recurrent layers, a critical architectural decision for video classification tasks is utilized. These layers excel at processing spatiotemporal information within video sequences. We configure the layer with parameters such as the number of filters, kernel size, and activation function to facilitate convolutional operations. The resulting sequences are subsequently processed through various other function layers, reducing frame dimensions to alleviate computational load, and Dropout layers, mitigating overfitting risks. The architecture is intentionally kept simple with a limited number of trainable parameters, commensurate with the scale of the dataset. A vital element is the incorporation of a final Dense layer with softmax activation, yielding probability distributions across action categories.

The constructed model is then compiled using categorical cross-entropy as the loss function, the Adam optimizer, and accuracy as the metric for evaluation. Training is initiated, incorporating an early stopping callback to prevent overfitting. This structure forms a cohesive and efficient framework for the in-depth analysis of spatiotemporal patterns within table tennis stroke videos. The model's adherence to best practices in architectural design and training strategies enhances its adaptability and potential for robust performance in action recognition tasks.

4 RESULTS AND ANALYSIS

The accuracy of the model was updated after every layer of training, and the results demonstrate a high level of accuracy. The training data accuracy reached a peak of 97%, while the validation accuracy reached 98.8%. Several factors contributed to these results.

Firstly, the volume and distribution of the data had a significant impact on the accuracy. Additionally, the image height, width, and sequence length all had a significant effect on the results, with accuracy ranging from 0.7408 for an image of dimensions 90*80 and a sequence length of 60, to 0.9876 for an image of dimensions 64*64 and a sequence length of 60.

RUN	Hand	Serve	Hand & Serve	Global
Run1	91.5	92.4	90.7	75.4
Run2	**93.2**	89.8	89.0	74.6
Run3	92.4	**94.1**	**92.4**	**81.4**
Run 4	89.0	89.8	86.4	72.0
Run 5	89.8	88.1	87.3	72.9

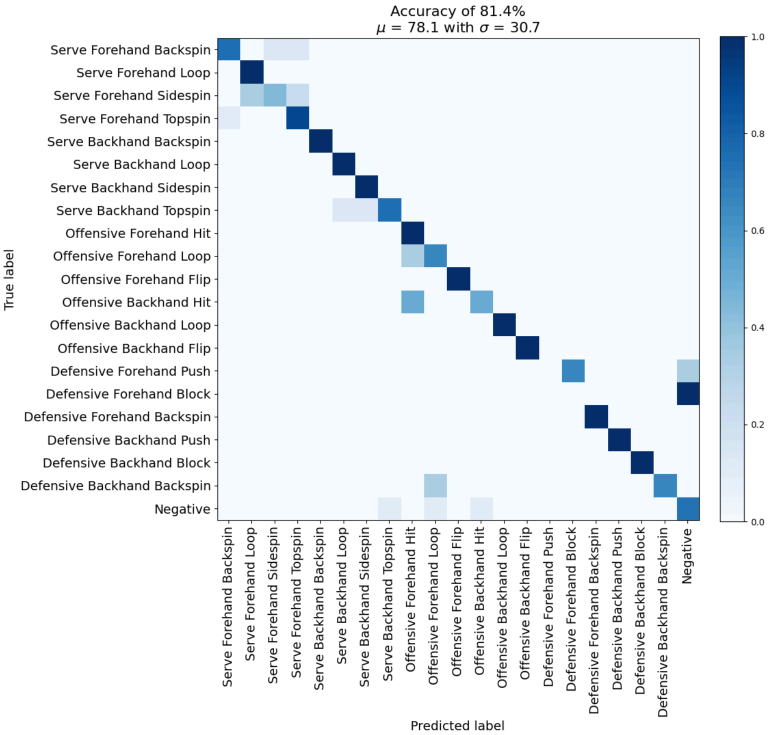
It is worth noting that the data distribution of some labels is highly biased towards certain classes, leading to biased learning. Over the course of five runs, the highest global accuracy achieved by the model was 81.4%.

5 DISCUSSION AND OUTLOOK

Throughout the training and validation phase, we have attained encouraging outcomes that lead us to conclude that overfitting is not present.

Nonetheless, the model's performance on the test data reveals that it has not effectively learned and is unable to generalize. In our opinion, this challenge can be remedied by augmenting the quantity of labeled data utilized in training.

Moreover, we posit that the low variability between the classes and the nature of the task contribute to this issue. Considering that a single class can be sampled in various ways for different players, such as right/left-handed or high/low experienced, we suggest that the dataset could be enhanced by increasing the coverage of the classes and reducing bias among them.



REFERENCES

[1] Pierre-Etienne Martin. Baseline Method for the Sport Task of MediaEval 2023 3D CNNs using Attention Mechanisms for Table Tennis Stoke Detection and Classification. *Working Notes Proceedings* of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online, 1-2 February 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[2] A. Erades, P. Martin, R. Vuillemot, B. Mansencal, R. Péteri, J. Morlier, S. Duffner, J. Benois-Pineau, Sportsvideo: A multimedia dataset for event and position detection in table tennis and swimming, *Working Notes Proceedings* of the MediaEval 2023 Workshop, Amsterdam, The Netherlands, 1-2 February 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[3] Kaustubh Milind Kulkarni, Sucheth Shenoy. Table Tennis Stroke Recognition Using Two-Dimensional Human Pose Estimation. *CVPR Sports Workshop 2021.*

[4] Kadir Aktas, Mehmet Demirel, Marilin Moor, Johanna Olesk, Gholamreza Anbarjafari.
Spatio-Temporal Based Table Tennis Hit Assessment Using LSTM Algorithm. *MediaEval'20, December 2020.*

[5] Anam Zahra, Pierre-Etienne Martin. Two Stream Network for Stroke Detection in Table Tennis. *MediaEval'21, December 2021.*

[6] R. Voeikov, N. Falaleev and R. Baikulov. TTNet: Real-time temporal and spatial video analysis of table tennis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2020 pp. 3866-3874. 10.1109/CVPRW50498.2020.00450

[7] Hai Nguyen-Truong, San Cao, N. A. Khoa Nguyen, Bang-Dang Pham, Hieu Dao, Minh-Quan Le, Hoang-Phuc Nguyen-Dinh, Hai-Dang Nguyen, Minh-Triet Tran. *HCMUS at MediaEval 2020:* Ensembles of Temporal Deep Neural Networks for Table Tennis Strokes Classification Task

[8] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Julien Morlier.
Siamese Spatio-Temporal Convolutional Neural Network for Stroke Classification in Table Tennis Games

[9] Zhou Jun. Biomechanical study of different techniques performed by elite athletes in table tennis. *Journal of Chemical and Pharmaceutical Research*, 6(2):589-591, 2014.

[10] S. S. Tabrizi, S. Pashazadeh, and V. Javani. 2020. Comparative Study of Table Tennis Forehand Strokes Classification Using Deep Learning and SVM. *IEEE Sensors Journal (2020), 1–1.*