# 3D attention mechanism for fine-grained classification of table tennis strokes using a Twin Spatio-Temporal Convolutional Neural Networks


**Pierre-Etienne Martin[1], Jenny Benois-Pineau[1], Renaud Péteri[2], Julien Morlier[3]**

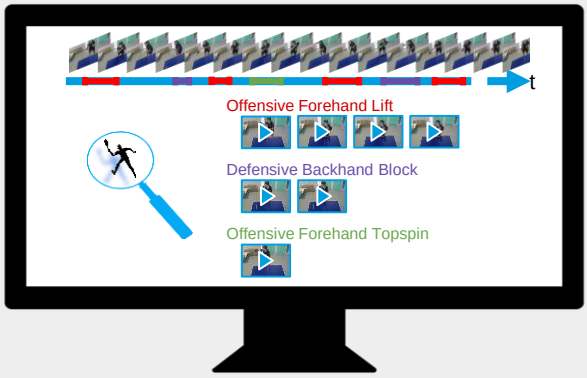
[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, Talence, France

[2]MIA, University of La Rochelle, La Rochelle, France

[3]IMS, University of Bordeaux, Talence, France


## Introduction

This paper tackles the problem of fine grained-action recognition from videos. We classify Table Tennis strokes in videos recorded in natural condition. The goal is to develop an interface where teachers and students can analyse their games for improving players performance. We introduce 3D attention blocks which are incorporated into a Twin network processing RGB and Optical Flow data in order to perform classification. The incorporated attention mechanism boosts both, convergence and classification performance.




**Interface to analyse players performance**

## TTStroke-21 Dataset

TTStroke-21 [1] is constituted of player-centred videos using GoPro cameras with 120 frames per second recorded in natural conditions. Experts in Table Tennis annotated the videos through a crowdsourced annotation platform using twenty stroke classes accordingly to the table tennis rules. A rejection class is built upon the filtered annotations.

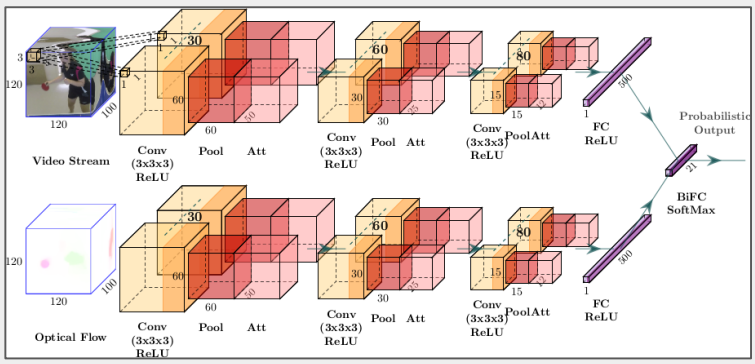
**Acquisition**

**Annotation platform**

**Samples TTStroke-21**

## Twin Spatio-Temporal Convolutional Neural Network with Attention Mechanism

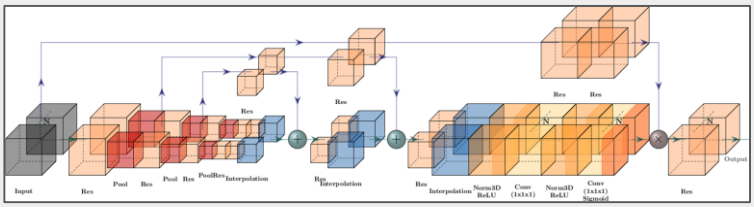
The Twin Spatio-Temporal Convolutional Neural Network - T-STCNN introduced in [1] is modified to incorporate attention mechanism. Attention blocks are inserted after the max-pooling layer. The model processes the video stream of size (W, H, T) = (120, 120, 100) and their estimated motion vectors $V=(v_x,v_y)$.




**T-STCNN architeture with attention mechanism**

## 3D Attention Blocks

The attention blocks were inspired by [1] and [3]. They are composed of 2 branches: the trunc branch (upper) and floating branch (lower). The floating branch is processed by several 3D Residual blocks and Max Pooling blocks, decreasing the feature map size and increasing the receptive field. Features of the lower level are then added to the upper level using 3D interpolations. The values of the floating branch are then mapped between 0 and 1 and multiply the features of the trunc branch, accentuating localized features.



**3D attention block architecture**

## Floating Branch Features Analysis

By analyzing the features of the floating branch output, we can determine where the model is focusing to perform classification. We can notice the difference of attention between the different blocks according to their position in the network. At the early stage, the focus is on the scene, such as the table. In the second stage, the body pat of the player are highlighted along with the border of the table, stressing the importance of the player position for classification. FInally, at the latest stage before feeding the feature to a fully connected layer, the racket and the ball are stressed: the focus is on the fine characteristics of the stroke.

**Attention 1** **Attention 2** **Attention 3**

**Soft mask branch output visualization for the different incorporated attention blocks**

## Classification results

Better classification results were obtained with the models using attention mechanisms. Also, the Twin models outperform the I3D models [4]. Also a faster convergence of the models using attention blocks were noticed.

**Table 1: Classification accuracy in % for the different tested models**

| Models | Train | Validation | Test |
|---|---|---|---|
| RGB-I3D [4] | 98 | 72.6 | 69.8 |
| RGB-STCNN [1] | 98.6 | 87 | 76.7 |
| RGB-STCNN with Attention | 96.9 | 88.3 | 85.6 |
| Flow-I3D [4] | 98.9 | 73.5 | 73.3 |
| Flow-STCNN [1] | 88.5 | 73.5 | 74.1 |
| Flow-STCNN with Attention | 96.4 | 83.5 | 79.7 |
| Two Stream-I3D [4] | 99.2 | 76.2 | 75.9 |
| Twin-STCNN [1] | 99 | 86.1 | 81.9 |
| **Twin-STCNN with Attention** | 97.3 | 87.8 | **87.3** |

## Discussion

In this work, we proposed a 3D attention mechanism through 3D attention blocks that can be translated to different models. We also offered an efficient method to train networks with different configurations. The attention mechanism efficiency was observed qualitatively through the appreciation of the highlighted features and the increased classification performance for the fine grained classification task. Application to such attention mechanism to different tasks and dataset are planned in order to assess better its transposition capacity.

## References


[1] P.-E. Martin et al., "Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks," in Multim. Tools Appl. 79, 27-28, 2020.

[2] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification", CVPR 2017, pp. 6450–6458.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", CoRR, vol. abs/1512.03385, 2015.

[4] J. Carreira and, A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," CoRR, vol. abs/1705.07750, 2017.