FINE-GRAINED ACTION DETECTION AND CLASSIFICATION IN TABLE TENNIS WITH SIAMESE SPATIO-TEMPORAL CONVOLUTIONAL NEURAL NETWORK

Pierre-Etienne Martin supervised by Jenny Benois-Pineau and Renaud Péteri

Univ. Bordeaux, LaBRI, France

INTRODUCTION

We detect and classify Table Tennis strokes in videos recorded in natural condition. The goal is to develop an intelligent computer environment where teachers and students can analyse their games for improving players performance.
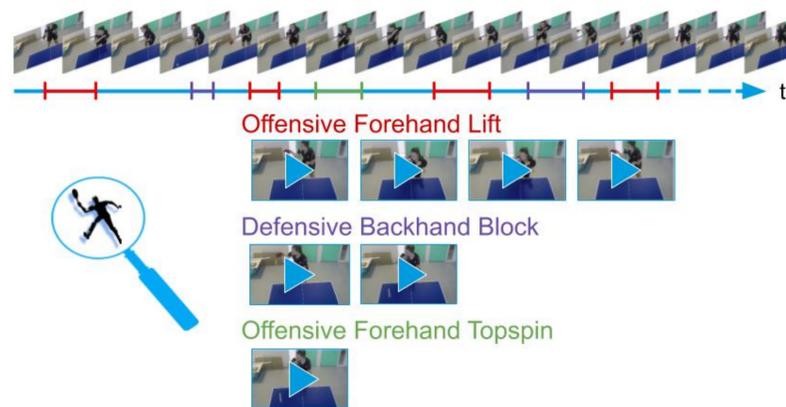

Figure 1: Intelligent computer framework proposal

TTSTROKE-21

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

Figure 2: Acquisition

Figure 3: TTStroke-21 samples

OPTICAL FLOW

From RGB video frames, motion vectors are computed[2]. A region of interest – ROI - of size (W, H) is then inferred from the center of mass of the foreground motion amplitude.

Figure 4: Optical Flow and spatial segmentation

SSTCNN

We introduced a Siamese Spatio-Temporal Convolutional Neural Network - SSTCNN to classify spatio-temporal data samples of size (W, H, T) = (120, 120, 100). The class is predicted from RGB video frames and their estimated motion vectors $V=(v_x, v_y)$.
For classification & detection task, we classify portions of video using a sliding window. Decision is made using a window of 150.
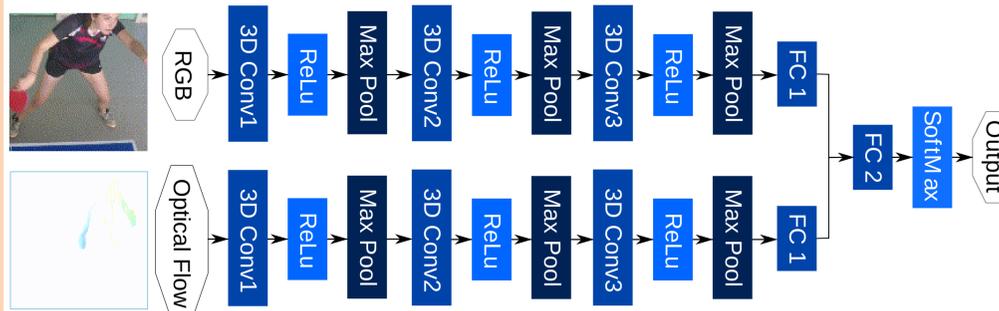

Figure 5: Siamese Architecture

RESULTS

Models were trained with data augmentation in spatial and temporal dimension using Stochastic Gradient descent with cross-entropy loss. Different normalization has been tested for the optical flow data. Best performances are obtained with our SSTCNN using mean-variance normalization. Strong overfitting is observed with the I3D model[3] trained from scratch.

Models	Accurcies		
	Train	Validation	Test
I3D - RGB	86.6	40	40.5
I3D - Flow	81.8	37.4	30.2
I3D - RGB + Flow	-	41.7	43.1
STCNN - RGB	97.4	88.7	81.9
STCNN - Flow	86.9	73.5	74.1
SSTCNN	99.1	90.4	**91.4**

CONCLUSION

- SSTCNN performs 91.4% of accuracy against 43.1% for the state of the art method[3]
- Dataset is still enriched
- Method is being improved to become robust to acquisition angles and obtain faster computation
- We reach 81.3% on detection & classification task

[1] P.-E. Martin et al., "Sport action recognition with siamese spatio-temporal cnns: Application to table tennis," in IEEE CBMI, 2018.
[2] P.-E. Martin et al., "Optimal choice of motion estimation methods for fine-grained action classification with 3D convolutional networks," in IEEE ICIP, 2019.
[3] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in IEEE CVPR, 2017.