Fine-Grained Action Detection and Classification in Table Tennis with Siamese Spatio-Temporal Convolutional Neural Network

Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri

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

Pierre-Etienne Martin[1] *supervised by Jenny Benois-Pineau*[1] *and Renaud Péteri*[2]

[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France
[2]MIA, La Rochelle University, La Rochelle, France

ABSTRACT

Human action recognition in videos is one of the key problems in visual data interpretation. Despite intensive research, the recognition of actions with low inter-class variability remains a challenge. To answer this problem, my thesis focus on fine-grained classification challenge using a Siamese Spatio-Temporal Convolutional Neural Network and apply it to a new dataset we have introduced `TTStroke-21`. Our model take as input data RGB images and Optical Flow and is able to reach an accuracy of 91.4% against 43.1% for our baseline on temporal segmented videos. Detection and classification in videos using a sliding temporal window leads to a score of 81.3% over the whole dataset.

***Index Terms*—** Deep Learning, Optical Flow, Action classification, Spatio-temporal convolution

1. INTRODUCTION

The target application of our research is fine grained action recognition in sports with the aim of improving athletes' performances. Without loss of generality, we are interested in recognition of strokes in table tennis. In `TTStroke-21`[1], twenty stroke classes and an additional rejection class are considered according to the rules of table tennis. This taxonomy was designed with professional table tennis teachers. We are working on videos recorded at the Faculty of Sports of the University of Bordeaux - STAPS. Students are the athletes filmed and the teachers supervise exercises conducted during the recording sessions. The recordings are markerless and allow players to perform in natural conditions. The goal is to develop an automatic analysis tool that teachers and students can use to analyse tennis table players games to improve their performances. This dataset is the first step of my thesis.

The second step is the classification process. A new Siamese Spatio-Temporal Convolutional Neural Network - SSTCNN - is introduced for this purpose [1]. Our model similarly processes RGB images and Optical Flow through a succession of spatio-temporal convolutions. A middle fusion is done before the calculation of the class scores. We compare the performances using our dataset with the baseline Two-Stream I3D method proposed in [2]. A temporal segmentation of table tennis strokes in videos is also be performed, based on temporal sliding windows and our SSTCNN classifier.

This work is supported by Region Nouvelle Aquitaine (grant CRISP) and Bordeaux Idex Initiative

2. TTSTROKE-21

`TTStroke-21` is constituted of player-centred videos using GoPro cameras with 120 frames per second recorded in natural conditions. Sequences have been recorded indoors using artificial light. Experts in Table Tennis annotate the videos through the annotation platform using twenty stroke classes accordingly to the table tennis rules. See Fig. 1 for an overview of the process. To obtain an exploitable dataset, annotations had to be processed by different filters to remove annotation errors and annotation where joined when part of videos have been annotated twice by different annotators. A total of 1058 annotations are then kept and a rejection class is built upon them.

3. METHOD

Classification of actions is performed for a single table tennis player performing a series of strokes. Full HD video frames are resized to 320×180 pixels and their Optical Flow (OF) is computed offline.

3.1. Optical Flow and Region of Interest

Different Optical flow methods were investigated [3]. A region of interest (ROI) of size (W, H) is then inferred from the center of mass of the foreground motion amplitude map (Fig. 2).

a. Video acquisition **b.** Annotation platform

Fig. 1. The `TTStroke-21` dataset overview

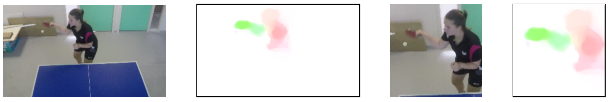

Fig. 2. ROI extraction based on optical flow

3.2. Architecture

Our SSTCNN is constituted of 2 individual branches with three 3D convolutional layers with 30, 60, 80 filter response maps, followed by a fully connected layer of size 500 (Fig. 3). One branch takes RGB values as input, and input of the other branch is the OF preliminary estimated on the current video frame. The 3D convolutional layers use $3 \times 3 \times 3$ space-time filters with a dense stride and padding set to 1 in each direction. The two branches are fused through a final fully connected layer of size 21 followed by a Softmax function for outputting a classification score.

3.3. Model Training

The '`Siamese`' model uses the full architecture presented in section 3.2 while '`RGB`' and '`Optical Flow`' models are constituted of one branch only. The optimization method is Stochastic Gradient Descent with Nesterov momentum.

3.4. Data Augmentation

Data augmentation is performed on the fly to save storage space. For spatial augmentation we apply random rotation, a random translation, and a random homothety both on RGB images and optical flow. Transformations are applied with respect to the center of the ROI. Finally we perform horizontal flip with probability of 0.5. For temporal augmentation we extract T successive frames following a normal distribution around the center of our stroke.

4. EXPERIMENTS AND RESULTS

To compare the performances of our models, we use the Two-Stream I3D model introduced by Carreira and Zisserman in [2] as our baseline and apply it to our dataset following their instructions for training (table 1). The RGB images and Optical Flow streams are trained separately and a late fusion by addition of the class scores is performed to classify the action. Our Siamese model perform the best.

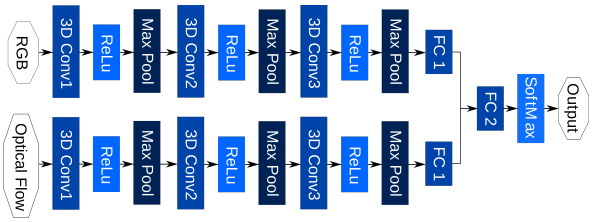

Fig. 3. SSTCNN - Siamese Spatio-Temporal Convolutional Neural Network

Table 1. Performance comparison of the different models

	Accuracies			
Models	**Val**	**Test**	**TestVote**	**TestAvg**
I3D (RGB)	40	40.5		
I3D (OptFlow)	37.4	30.2		
I3D (RGB + OptFlow)	41.7	43.1		
RGB	88.7	78.5	78.5	81.9
Optical Flow	47.8	44	44	44.8
Early Fusion (RGB + OptFlow)	84.4	73.3	74.1	75
Late Fusion (RGB + OptFlow)	62.2	57.7	59.5	70.7
Siamese	90.43	87.9	88.8	**91.4**

In recent work [3] i) we improve these results by changing the normalization method of the OF, ii) we perform both detection and classification simultaneously using a sliding temporal window on the whole dataset leading to a score of 81.3%.

5. CONCLUSION AND PERSPECTIVES

My thesis aims to improve athletes performances by developing new methods and tools for coaches and students. For now on, we are able to segment temporally and classify games of Table Tennis players. In a near future, other information will be extracted in the method framework, such as characterizing the quality of a performed stroke, or establishing player statistics. The same protocol could be extended to other sports by adapting classes and size of our video cuboids according to the specific rules of the sport of interest. However, it requires to have a dataset dedicated for each sport. In our case, the `TTStroke-21` dataset, is still enriched to improve the performances of our SSTCNN model and to make possible better tools for coaches and students.

6. REFERENCES

[1] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier, "Sport action recognition with siamese spatio-temporal cnns: Application to table tennis," in *CBMI 2018*. 2018, pp. 1–6, IEEE.

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

[3] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier, "Optimal choice of motion estimation methods for fine-grained action classification with 3d convolutional networks," in *ICIP*. 2019, IEEE.

[4] Philippe Weinzaepfel, Jérôme Revaud, Zaïd Harchaoui, and Cordelia Schmid, "Deepflow: Large displacement optical flow with deep matching," in *IEEE ICCV*, 2013, pp. 1385–1392.