Sports Video Annotation: Detection of Strokes in Table Tennis task for MediaEval 2019

Pierre-Etienne Martin[1], Jenny Benois-Pineau[1], Boris Mansencal[1],
Renaud Péteri[2], Laurent Mascarilla[2], Jordan Calandre[2],
Julien Morlier[3]
[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France
[2]MIA, La Rochelle University, La Rochelle, France
[3]IMS, University of Bordeaux, Talence, France
pierre-etienne.martin@u-bordeaux.fr,jenny.benois-pineau@u-bordeaux.fr,boris.mansencal@labri.fr
renaud.peteri@univ-lr.fr,lmascari@univ-lr.fr,jordan.calandre1@univ-lr.fr
julien.morlier@u-bordeaux.fr

ABSTRACT

Action detection and classification is one of the main challenges in visual content analysis and mining. Sport video analysis has been a very popular research topic, due to the variety of application areas, ranging from multimedia intelligent devices with user-tailored digests, up to analysis of athletes' performances. Datasets with sport activities are available now for benchmarking of methods. A large amount of work is also devoted to the analysis of sport gestures using motion capture systems. However, body-worn sensors and markers could disturb the natural behaviour of sports players. Furthermore, motion capture devices are not always available for potential users, be it a University Faculty or a local sport team. Coming years will build upon the basic "Sports Video Annotation: Detection of Strokes in Table Tennis" task offered in 2019. The ultimate goal of this research is to produce automatic annotation tools for sport faculties, local clubs and associations to help coaches to better assess and advise athletes during training

1 INTRODUCTION

Action detection and classification is one of the main challenges in visual content analysis and mining [11]. Sport video analysis has been a very popular research topic, due to the variety of application areas, ranging from multimedia intelligent devices with user-tailored digests, up to analysis of athletes' performances[2].The Sport Video Annotation project was initiated between the Faculty of Sports STAPS of the University of Bordeaux, the LaBRI - Université de Bordeaux and the MIA lab. - La Rochelle University. It is supported by the CNRS federation MIRES and the New Aquitaine Region in the framework of an "APP Recherche". The goal of this project is to develop artificial intelligence and multimedia indexing methods for the recognition of table tennis sports activities. The aim is to evaluate the performance of athletes, with a particular focus on students, in order to develop optimal training strategies. To that aim, a video corpus named `TTStroke-21` was recorded with volunteered players. These data represent a large scientific interest for the Multimedia community participating in the MediaEval campaign.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

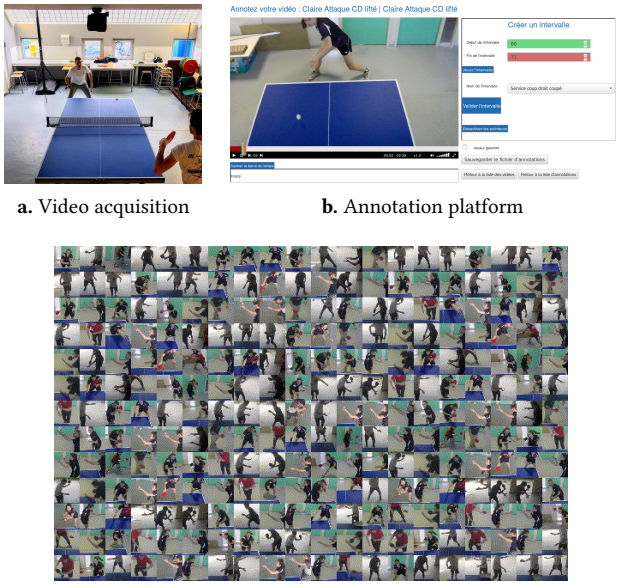
Other datasets such as UCF-101 [10], HMDB [5], [4] and AVA [3] are used as benchmarks for action classification methods. Others, such as the Olympic Sports dataset [9] focus on sport actions only. However none of them is dedicated to a specific sport and its associated rules. Furthermore, `TTStroke-21` is annotated manually by professional players or teachers of Table Tennis, making the annotation process longer, but more temporally and qualitatively accurate. Classification methods as I3D model [1] or LTC model [12] performing well on UCF-101 dataset inspired the work done in [7] and [8] through a SSTCNN - Siamese Spatio Temporal Convolutional Neural Network. Here the video stream and derived computed optical flow are passed through the branches of the SSTCNN. The similarity of actions - strokes - in `TTStroke-21` makes the classification task challenging and the multi-modal method seemed to improve performances. In [6], spatio temporal dependencies are learned from the video using only RGB images and scores are promising but are still below the multi-modal methods of I3D.

2 PARTICULAR CONDITIONS

Because `TTStroke-21` is constituted of videos with identifiable players of Table Tennis, this dataset is subject to particular conditions in order to respect the personal data and privacy of the players. These Special Conditions apply to the use of Images generated in the framework of the program Sports video annotations: classification of strokes in table tennis, for the implementation of the MediaEval program. They constitute the specific usage agreement referred to in the Usage agreement for the MediaEval 2019 Research Collections, signed between the User and the University of Delft. The full and complete acceptance, without any reservation, of these Special Conditions is a mandatory prerequisite for the provision of the Images as part of the MediaEval programme. A complete reading of these conditions are necessary and engage the user, for example, to obscure the faces (blurring, black banner, etc.) before any publication and to destroy the data by October 1st 2020.

3 DATASET DESCRIPTION

In MediaEval 2019, we deliver a subset of `TTStroke-21` data set which has been specifically recorded in a sport faculty facility using a light-weight equipment, such as GoPro cameras. It is constituted of player-centred videos recorded in natural conditions without markers or sensors, see Fig 1. It comprises 20 table tennis strokes



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

c. Dataset samples

Figure 1: `TTStroke-21`

classes, i.e. 8 services, 6 offensive strokes and 6 defensive strokes. This taxonomy was designed with professional table tennis teachers.
All videos are recorded in MP4 format.

The organisation of the delivered data is as follows:

- The provided dataset is split into two subsets: i)training set and ii)test set;
- In each directory, there are several videos (in MP4 format) and each video may contain several actions;
- Each video file is accompanied with a XML file describing the actions present in the video;
- For each action there are 3 attributes: the starting frame, the ending frame, and the stroke class;
- In the train set XML files, all the attributes are specified but in the test set XML files, only the starting and ending frames are specified while the stroke class attribute is purposely set to an invalid value ("Unknown") and should be updated by the participants to one of the 20 valid classes.

.

4 TASK DESCRIPTION

The Sport Video Annotation task consists in assigning a label from a given taxonomy of 20 classes of Table Tennis strokes to each action delimited by starting frame and ending frame in each test video file.
Participants may submit up to four runs. For each runs, they must provide one XML file per video file, with the actions associated to the recognised stroke class. Runs may be submitted as an archive (zip or tar.gz file) with each run in a different directory. Participants should also indicate if any external data (other dataset, pretrained networks, ...) was used to compute their runs. The task is considered fully automatic. Once the video are provided to the system, results should be produced without any human intervention.

5 EVALUATION

In MediaEval 2019 we propose a light-weight classification task. It consists in classification of table tennis strokes which temporal borders are supplied in the XML files accompanying each video file. Hence for each test video the participants are invited to produce an xml file in which each stroke is labelled accordingly to the given taxonomy. This means that the label "unknown" has to be replaced by the label of the stroke class which participant's system has assigned. All submissions will be evaluated in terms of `per-class accuracy`(PCA) and of global accuracy (GA).The PCA is computed for each i-th class as:

$$PCA_i = TP_i/(N_{gti}) \quad (1)$$

Here TP_i is the number of True Positives, i.e. correctly labelled, by the participant's system, strokes for the given i-th class,N_{gti} is the number of recorded strokes of the i-th class in the test dataset.

$$GA = TP/(N_{gt}) \quad (2)$$

Here $TP = \sum TP_i$ is the number of correctly labelled strokes for the whole dataset, and N_{gt} is the number of strokes in the ground truth - the whole test set.

6 DISCUSSION

Participants are welcome to share their difficulties and the results even if they are negative. Better understanding of automatic classification methods are easier when all aspects of the methods are shared.

Thank you for participating at MediaEval 2019 and more specifically to our task: "Sports Video Annotation: Detection of Strokes in Table Tennis". We look forward to seeing you at the MediaEval Workshop and to discussing your results further.

ACKNOWLEDGMENTS

We would like to thank all the players and annotators who contributed to `TTStroke-21`, Alain Coupet for his dedication to the project, Xavier Daverat and Chantal Durand for their help on the Particular Conditions formulation.
This work was supported by Region of Nouvelle Aquitaine grant CRISP and Bordeaux Idex Initiative.

REFERENCES

[1] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *CoRR* abs/1705.07750 (2017). arXiv:1705.07750

[2] Moritz Einfalt, Dan Zecha, and Rainer Lienhart. 2018. Activity-Conditioned Continuous Human Pose Estimation for Performance Analysis of Athletes Using the Example of Swimming. In *WACV*. 446–455.

[3] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2017. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. *CoRR* abs/1705.08421 (2017).

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back,

Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017).

[5] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. 2011. HMDB: A large video database for human motion recognition. In *ICCV*. IEEE Computer Society, 2556–2563.

[6] Zheng Liu and Haifeng Hu. 2019. Spatiotemporal Relation Networks for Video Action Recognition. *IEEE Access* 7 (2019), 14969–14976.

[7] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2018. Sport Action Recognition with Siamese Spatio-Temporal CNNs: Application to Table Tennis. In *CBMI 2018*. IEEE, 1–6.

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

[9] Juan Carlos Niebles, Chih-Wei Chen, and Fei-Fei Li. 2010. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In *ECCV 2010*. 392–405.

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR* 1212.0402 (2012). arXiv:1212.0402

[11] Andrei Stoian, Marin Ferecatu, Jenny Benois-Pineau, and Michel Crucianu. 2016. Fast Action Localization in Large-Scale Video Archives. *IEEE Trans. Circuits Syst. Video Techn.* 26, 10 (2016), 1917–1930.

[12] Gül Varol, Ivan Laptev, and Cordelia Schmid. 2018. Long-Term Temporal Convolutions for Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1510–1517.