Sports Video Classification: Classification of Strokes in Table Tennis for MediaEval 2020

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, Talence, France
[2]MIA, La Rochelle University, La Rochelle, France
[3]IMS, University of Bordeaux, Talence, France
mediaeval.sport.task@diff.u-bordeaux.fr

ABSTRACT

Fine grained action classification has raised new challenges compared to classical action classification problem. Sport video analysis is 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. Running since 2019 as a part of MediaEval, we offer a task which consists in classifying table tennis strokes from videos recorded in natural conditions at the University of Bordeaux. The aim is to build tools for teachers, coaches and players to analyse table tennis games. Such tools could lead to an automatic profiling of the player and the training session could then be adapted for improving sports skills more efficiently.

1 INTRODUCTION

Action detection and classification is one of the main challenges in visual content analysis and mining [26]. Over the last few years, the number of datasets for action classification has drastically increased in terms of video content, resolution, localization and number of classes. However the latest research shows that classification performed using deep neural networks often focuses on the whole scene and the background and not on the action itself.

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' performance [5]. The Sport Video Classification project was initiated by the Faculty of Sports (STAPS) and the computer science laboratories (LaBRI) of the University of Bordeaux, and by the MIA laboratory of La Rochelle University[1]. The goal of this project is to develop artificial intelligence and multimedia indexing methods for the recognition of table tennis sport activities. The ultimate goal 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 volunteer players. These data are of great scientific interest for the Multimedia community participating in the MediaEval campaign.

[1]This work was supported by the New Aquitania Region through CRISP project - ComputeR vIsion for Sport Performance and the MIRES federation.

MediaEval'20, 14-15 December 2020, Online

Several datasets such as UCF-101 [24], HMDB [10] and AVA [7] have been used for many years as benchmarks for action classification methods. In [15], spatio-temporal dependencies are learned from the video using only RGB images for classification. This method is promising but its scores are still below multi-modal methods such I3D [4]. More recently, datasets have been enriched, like JHMDB [8] and Kinetics [2, 3, 9] or fused like AVA_Kinetics [12]. Some also focus on the intra-class dissimilarity such as the Something-Something dataset. Others, such as the Olympic Sports dataset [22], focus on sport actions only. However those datasets are not dedicated to a specific sport and its associated rules. Few datasets focus on fine-grained classification. We can cite FineGym [23], introduced recently, which focuses on gymnastic videos, and our TTStroke21 [21] comprising table tennis strokes.

`TTStroke-21` is manually annotated by professional players or teachers of table tennis, making the annotation process more time consuming, but more temporally and qualitatively accurate. Classification methods such as I3D model [4] or LTC model [28] performing well on UCF-101 dataset inspired the work done in [18, 21] which introduces a TSTCNN - Twin Spatio Temporal Convolutional Neural Network. Here, the video stream and derived computed optical flow are passed through the branches of the TSTCNN. In [19] the normalization of the flow is also investigated to improve the classification score while in [20] an attention block is introduced to improve the performances and speed of convergence. The inter-similarity of actions - strokes - in `TTStroke-21` makes the classification task challenging and the multi-modal method seemed to improve performances. To better understand learned features and classification process taking place in the TSTCNN, we also developed a new visualization technique [6].

Recent work focusing on table tennis [30] tries to get the tactics of the players based on their performance during matches using a Markov chain model. In [14, 27, 32] stroke recognition is performed using sensors. In [29] segmentation of the player, ball coordinates, event detection is explored while [13, 31] focus solely on the trajectory of the ball.

In this task overview paper, in section 2, we introduce the specific conditions of usage of this data, then describe `TTStroke-21` and the task respectively in sections 3 and 4. The evaluation method is explained in section 5. Supplementary notes are shared in section 6. More information can be found on the dedicated GitHub web page[2].

[2]https://multimediaeval.github.io/2020-Sports-Video-Classification-Task/

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

Figure 1: TTStroke-21 acquisition process

2 SPECIFIC CONDITIONS OF USAGE

TTStroke-21 is constituted of videos with players playing table tennis in natural conditions. Even if we are using an automatic tool for blurring players' faces, some faces are misdetected on few frames and thus some players remain identifiable. In order to respect the personal data and privacy of the players, this dataset is subject to a usage agreement, referred to as *Special Conditions*. These *Special Conditions* apply to the use of videos, referred to as Images, generated in the framework of the program Sports video classification: classification of strokes in table tennis, for the implementation of the MediaEval program. They correspond to the specific usage agreement referred to in the *Usage agreement for the MediaEval 2020 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 2020 evaluation campaign. A complete reading of these conditions is necessary and requires the user, for example, to obscure the faces (blurring, black banner, etc.) in the video before use in any publication and to destroy the data by October 1st, 2021.

3 DATASET DESCRIPTION

In the MediaEval 2020 campaign, we release the same subset of the TTStroke-21 dataset than last year. The only difference is the blurring of the faces and the specification if the player is right-handed or left-handed. The dataset has been 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 stroke classes, i.e. 8 services: `Serve Forehand Backspin`, `Serve Forehand Loop`, `Serve Forehand Sidespin`, `Serve Forehand Topspin`, `Serve Backhand Backspin`, `Serve Backhand Loop`, `Serve Backhand Sidespin`, `Serve Backhand Topspin`; 6 offensive strokes: `Offensive Forehand Hit`, `Offensive Forehand Loop`, `Offensive Forehand Flip`, `Offensive Backhand Hit`, `Offensive Backhand Loop`, `Offensive Backhand Flip`; and 6 defensive strokes: `Defensive Forehand Push`, `Defensive Forehand Block`, `Defensive Forehand Backspin`, `Defensive Backhand Push`, `Defensive Backhand Block`, `Defensive Backhand Backspin`. Also, all the strokes can be divided in two super-classes: Forehand and Backhand. This taxonomy was designed with professional table tennis teachers.

All videos are recorded in MPEG-4 format. Unlike the task at MediaEval 2019 [16], most of the faces are blurred. To do so, faces are detected with OpenCV deep learning face detector, based on the Single Shot Detector (SSD) framework with a ResNet base network, for each frame of the original video. The detected face is blurred and frames are re-encoded in a video.

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 MPEG-4 format) and each video may contain several actions;
- Each video file is provided with a XML file describing the actions present in the video and if the player is right-handed or left-handed;
- Each action has 3 attributes: the starting frame, the ending frame, and the stroke class;
- In the train set XML files, all the attributes are specified. In the test set XML files, only the starting and ending frames are specified. The stroke class attribute is purposely set to 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, for each action of each test video, in assigning a label using a given taxonomy of 20 classes of table tennis strokes.

Participants may submit up to five runs. For each run, they must provide one XML file per video file containing, with the actions associated with 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, such as other dataset or 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

For MediaEval 2020, 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 default label "unknown" has to be replaced by the label of the stroke class that the participant's system has assigned. All submissions will be evaluated in terms of *per-class accuracy* (A_i) and of global accuracy (GA).

The organizers will also provide to the participants different confusion matrices: one considering all the classes, and others considering the type of the stroke such as: offensive, defensive and defensive and/or using forehand and backhand superclasses of the strokes.

6 DISCUSSION

The participants from last years have reached a maximum accuracy of 22.9% [17], 14.1%[1] and 11.3% [25] leaving room for improvement. Participants are welcome to share their difficulties and their results even if they seem not sufficiently good.

REFERENCES

[1] Jordan Calandre, Renaud Péteri, and Laurent Mascarilla. 2019. Optical Flow Singularities for Sports Video Annotation: Detection of Strokes in Table Tennis, See [11].

[2] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. *CoRR* abs/1808.01340 (2018).

[3] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A Short Note on the Kinetics-700 Human Action Dataset. *CoRR* abs/1907.06987 (2019).

[4] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. (2017), 4724–4733.

[5] 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 *IEEE WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. 446–455.

[6] Kazi Ahmed Asif Fuad, Pierre-Etienne Martin, Romai Giot, Romain Bourqui, Jenny Benois-Pineau, and Akka Zemmari. 2020. Feature Understanding in 3D CNNs for Actions Recognition in Video. In *Tenth International Conference on Image Processing Theory, Tools and Applications, IPTA 2020, Paris, France, November 9-12, 2020*. 1–6.

[7] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. (2018), 6047–6056.

[8] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. 2013. Towards Understanding Action Recognition. In *IEEE ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 3192–3199.

[9] Will Kay, João 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).

[10] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. 2011. HMDB: A large video database for human motion recognition. In *IEEE ICCV 2011, Barcelona, Spain, November 6-13, 2011*, Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool (Eds.). IEEE Computer Society, 2556–2563.

[11] Martha A. Larson, Steven Alexander Hicks, Mihai Gabriel Constantin, Benjamin Bischke, Alastair Porter, Peijian Zhao, Mathias Lux, Laura Cabrera Quiros, Jordan Calandre, and Gareth Jones (Eds.). 2020. *Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019*. CEUR Workshop Proceedings, Vol. 2670. CEUR-WS.org.

[12] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The AVA-Kinetics Localized Human Actions Video Dataset. *CoRR* abs/2005.00214 (2020).

[13] Hsien-I Lin, Zhangguo Yu, and Yi-Chen Huang. 2020. Ball Tracking and Trajectory Prediction for Table-Tennis Robots. *Sensors* 20, 2 (2020).

[14] Ruichen Liu, Zhelong Wang, Xin Shi, Hongyu Zhao, Sen Qiu, Jie Li, and Ning Yang. 2019. Table Tennis Stroke Recognition Based on Body Sensor Network. In *IDCS 2019, Naples, Italy, October 10-12, 2019, Proceedings (Lecture Notes in Computer Science)*, Raffaele Montella, Angelo Ciaramella, Giancarlo Fortino, Antonio Guerrieri, and Antonio Liotta (Eds.), Vol. 11874. Springer, 1–10.

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

[16] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Laurent Mascarilla, Jordan Calandre, and Julien Morlier. 2019. Sports Video Annotation: Detection of Strokes in Table Tennis Task for MediaEval 2019, See [11].

[17] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, and Julien Morlier. 2019. Siamese Spatio-Temporal Convolutional Neural Network for Stroke Classification in Table Tennis Games, See [11].

[18] 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, La Rochelle, France, September 4-6, 2018*. IEEE, 1–6.

[19] 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 *IEEE ICIP 2019, Taipei, Taiwan, September 22-25, 2019*. IEEE, 554–558.

[20] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. 3D attention mechanisms in Twin Spatio-Temporal Convolutional Neural Networks. Application to action classification in videos of table tennis games.. In *2ICPR2020 - MiCo Milano Congress Center, Italy, 10-15 January 2021*.

[21] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks. *Multim. Tools Appl.* 79, 27-28 (2020), 20429–20447.

[22] Juan Carlos Niebles, Chih-Wei Chen, and Fei-Fei Li. 2010. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In *Computer Vision - ECCV 2010, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II (Lecture Notes in Computer Science)*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.), Vol. 6312. Springer, 392–405.

[23] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. *CoRR* abs/2004.06704 (2020).

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

[25] Siddharth Sriraman, Srinath Srinivasan, Vishnu K. Krishnan, Bhuvana J, and T. T. Mirnalinee. 2019. MediaEval 2019: LRCNs for Stroke Detection in Table Tennis, See [11].

[26] 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.

[27] 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.

[28] 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.

[29] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. *CoRR* abs/2004.09927 (2020).

[30] Jiachen Wang, Kejian Zhao, Dazhen Deng, Anqi Cao, Xiao Xie, Zheng Zhou, Hui Zhang, and Yingcai Wu. 2020. Tac-Simur: Tactic-based Simulative Visual Analytics of Table Tennis. *IEEE Trans. Vis. Comput. Graph.* 26, 1 (2020), 407–417.

[31] Erwin Wu and Hideki Koike. 2020. FuturePong: Real-time Table Tennis Trajectory Forecasting using Pose Prediction Network. In *CHI 2020, Honolulu, HI, USA*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–8.

[32] Kun Xia, Hanyu Wang, Menghan Xu, Zheng Li, Sheng He, and Yusong Tang. 2020. Racquet Sports Recognition Using a Hybrid Clustering Model Learned from Integrated Wearable Sensor. *Sensors* 20, 6 (2020), 1638.