Sports Video: Fine-Grained Action Detection and Classification of Table Tennis Strokes from Videos for MediaEval 2021

Pierre-Etienne Martin[1], Jordan Calandre[2], Boris Mansencal[3], Jenny Benois-Pineau[3], Renaud Péteri[2], Laurent Mascarilla[2], Julien Morlier[4]
[1]CCP Department, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany
[2]MIA, La Rochelle University, La Rochelle, France
[3]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, Talence, France
[4]IMS, University of Bordeaux, Talence, France
mediaeval.sport.task@diff.u-bordeaux.fr

ABSTRACT

Sports video analysis is a prevalent research topic due to the variety of application areas, ranging from multimedia intelligent devices with user-tailored digests up to analysis of athletes' performance. The Sports Video task is part of the MediaEval 2021 benchmark. This task tackles fine-grained action detection and classification from videos. The focus is on recordings of table tennis games. Running since 2019, the task has offered a classification challenge from untrimmed video recorded in natural conditions with known temporal boundaries for each stroke. This year, the dataset is extended and offers, in addition, a detection challenge from untrimmed videos without annotations. This work aims at creating tools for sports coaches and players in order to analyze sports performance. Movement analysis and player profiling may be built upon such technology to enrich the training experience of athletes and improve their performance.

1 INTRODUCTION

Action detection and classification are one of the main challenges in computer vision [20]. Over the last few years, the number of datasets and their complexity dedicated to action classification has drastically increased [12]. Sports video analysis is one branch of computer vision and applications in this area range from multimedia intelligent devices with user-tailored digests, up to analysis of athletes' performance [4, 21, 28]. A large amount of work is devoted to the analysis of sports gestures using motion capture systems. However, body-worn sensors and markers could disturb the natural behavior of sports players. This issue motivates the development of methods for game analysis using non-invasive equipment such as video recordings from cameras.

The Sports Video Classification project was initiated by the Sports Faculty (STAPS) and the computer science laboratory LaBRI of the University of Bordeaux, and the MIA laboratory of La Rochelle University[1]. This project aims to develop artificial intelligence and multimedia indexing methods for the recognition of table tennis activities. The ultimate goal is to evaluate the performance of athletes, with a particular focus on students, to develop optimal training strategies. To that aim, the video corpus named `TTStroke-21` was recorded with volunteer players.

Datasets such as UCF-101 [26], HMDB [6, 8], AVA [5] and Kinetics [1, 2, 7, 9, 25] are being use in the scope of action recognition with, year after year, an increasing number of considered videos and classes. Few datasets focus on fine-grained classification in sports such as FineGym [24] and TTStroke21 [19].

To tackle the increasing complexity of the datasets, we have on one hand methods getting the most of the temporal information: for example, in [11], where spatio-temporal dependencies are learned from the video using only RGB data. And on the other hand, methods combining other modalities extracted from videos, such as the optical flow [3, 18, 27]. The inter-similarity of actions - strokes - in `TTStroke-21` makes the classification task challenging, and both cited aspects shall be used to improve performance.

The following sections present the Sport task this year and its specific terms of use. Complementary information on the task may be found on the dedicated page from the MediaEval website[2].

2 TASK DESCRIPTION

This task uses the `TTStroke-21` database [19]. This dataset is constituted of recordings of table tennis players performing in natural conditions. This task offers researchers an opportunity to test their fine-grained classification methods for detecting and classifying strokes in table tennis videos. Compared to the Sports Video 2020 edition, this year, we extend the task with detection, and enrich the data set with new and more diverse stroke samples. The task now offers two subtasks. Each subtask has its own split of the dataset, leading to different train, validation, and test sets.

Participants can choose to participate in only one or both subtasks and submit up to five runs for each. The participants must provide one XML file per video file present in the test set for each run. The content of the XML file varies according to the subtask. Runs may be submitted as an archive (zip file), with each run in a different directory for each subtask. Participants should also submit a working notes paper, which describes their method and indicates if any external data, such as other datasets or pretrained networks, was used to compute their runs. The task is considered fully automatic: once the videos are provided to the system, results should be produced without human intervention. Participants are encouraged to release their code publicly with their submission. This year, a baseline for both subtasks was shared publicly [13].

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

MediaEval'21, December 13-15 2021, Online

[2]https://multimediaeval.github.io/editions/2021/tasks/sportsvideo/

2.1 Subtask 1 - Stroke Detection

Participants must build a system that detects whether a stroke has been performed, whatever its class, and extract its temporal boundaries. The aim is to distinguish between moments of interest in a game (players performing strokes) from irrelevant moments (time between strokes, picking up the ball, having a break…). This subtask can be a preliminary step for later recognizing a stroke that has been performed.

Participants have to segment regions where a stroke is performed in the provided videos. Provided XML files contain the stroke temporal boundaries (frame index of the videos) related to the train and validation sets. We invite the participants to fill an XML file for each test video in which each stroke should be temporally segmented frame-wise following the same structure.

For this subtask, the videos are not shared across train, validation, and test sets; however, a same player may appear in the different sets. The Intersection over Union (IoU) and Average Precision (AP) metrics will be used for evaluation. Both are usually used for image segmentation but are adapted for this task:

- **Global IoU:** the frame-wise overlap between the ground truth and the predicted strokes across all the videos.
- **Instance AP:** each stroke represents an instance to be detected. Detection is considered True when the IoU between prediction and ground truth is above an IoU threshold. 20 thresholds from 0.5 to 0.95 with a step of 0.05 are considered, similarly to the COCO challenge [10]. This metric will be used for the final ranking of participants.

2.2 Subtask 2 - Stroke Classification

This subtask is similar to the main task of the previous edition [15]. This year the dataset is extended, and a validation set is provided.

Participants are required to build a classification system that automatically labels video segments according to a performed stroke. There are 20 possible stroke classes. The temporal boundaries of each stroke are supplied in the XML files accompanying each video in each set. The XML files dedicated to the train and validation sets contain the stroke class as a label, while in the test set, the label is set to "Unknown". Hence for each XML file in the test set, the participants are invited to replace the default label "Unknown" by the stroke class that the participant's system has assigned according to the given taxonomy.

For this subtask, the videos are shared across the sets following a random distribution of all the strokes with the proportions of 60%, 20% and 20% respectively for the train, validation and test sets. All submissions will be evaluated in terms of global accuracy for ranking and detailed with per-class accuracy.

Last year, the best global accuracy (31.4%) was obtained by [22] using Channel-Separated CNN. [17] is second (26.6%) using 3D attention mechanism and [23] third (16.7%) using pose information and cascade labelling method. Improvement has been observed compared to the previous edition [14] with a best accuracy of 22.9% [16]. This improvement seems to be correlated by various factors such as: i) multi-modal methods, ii) deeper and more complex CNN capturing simultaneously spatial and temporal features, and iii) class decision following a cascade method.

Figure 1: Key frames of a same stroke from TTStroke-21

3 DATASET DESCRIPTION

The dataset has been recorded at the STAPS using lightweight equipment. It is constituted of player-centered videos recorded in natural conditions without markers or sensors, see Fig 1. Professional table tennis teachers designed a dedicated taxonomy. The dataset comprises 20 table tennis stroke classes: height services, six offensive strokes, and six defensive strokes. The strokes may be divided in two super-classes: Forehand and Backhand.

All videos are recorded in MPEG-4 format. We blurred fhe faces of the players for each original video frame using OpenCV deep learning face detector, based on the Single Shot Detector (SSD) framework with a ResNet base network. A tracking method has been implemented to decrease the false positive rate. The detected faces are blurred, and the video is re-encoded in MPEG-4.

Compared with Sports Video 2020 edition, this year, the data set is enriched with new and more diverse video samples. A total of 100 minutes of table tennis games across 28 videos recorded at 120 frames per second is considered. It represents more than 718 000 frames in HD (1920 × 1080). An additional validation set is also provided for better comparison across participants. This set may be used for training when submitting the test set's results. Twenty-two videos are used for the Stroke Classification subtask, representing 1017 strokes randomly distributed in the different sets following the previously given proportions. The same videos are used in the train and validation sets of the Segmentation subtask, and six additional videos, without annotations, are dedicated to its test set.

4 SPECIFIC TERMS OF USE

Although faces are automatically blurred to preserve anonymity, some faces are misdetected, and thus some players remain identifiable. In order to respect the personal data of the players, this dataset is subject to a usage agreement, referred to as *Special Conditions*. The complete acceptance of these *Special Conditions* is a mandatory prerequisite for the provision of the Images as part of the MediaEval 2021 evaluation campaign. A complete reading of these conditions is necessary and requires the user, for example, to obscure the faces (blurring, black banner) in the video before use in any publication and to destroy the data by October 1st, 2022.

5 DISCUSSIONS

This year the Sports Video task of MediaEval proposes two subtasks: i) Detection and ii) Classification of strokes from videos. Even if the players' faces are blurred, the provided videos still fall under particular usage conditions that the participants need to accept. Participants are encouraged to share their difficulties and their results even if they seem not sufficiently good. All the investigations, even when not successful, may inspire future methods.

ACKNOWLEDGMENTS

Many thanks to the players, coaches, and annotators who contributed to TTStroke-21.

REFERENCES

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

[2] 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).

[3] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*. IEEE Computer Society, 4724–4733.

[4] 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.

[5] 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.

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

[7] 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).

[8] 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.

[9] 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).

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8693. Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

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

[12] Pierre-Etienne Martin. 2020. *Fine-Grained Action Detection and Classification from Videos with Spatio-Temporal Convolutional Neural Networks. Application to Table Tennis. (Détection et classification fines d'actions à partir de vidéos par réseaux de neurones à convolutions spatio-temporelles. Application au tennis de table).* Ph.D. Dissertation. University of La Rochelle, France. https://tel.archives-ouvertes.fr/tel-03128769

[13] Pierre-Etienne Martin. 2021. Spatio-Temporal CNN baseline method for the Sports Video Task of MediaEval 2021 benchmark. In *MediaEval (CEUR Workshop Proceedings)*. CEUR-WS.org.

[14] 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. In *MediaEval (CEUR Workshop Proceedings)*, Vol. 2670. CEUR-WS.org.

[15] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Laurent Mascarilla, Jordan Calandre, and Julien Morlier. 2020. Sports Video Classification: Classification of Strokes in Table Tennis for MediaEval 2020. In *MediaEval (CEUR Workshop Proceedings)*, Vol. 2882. CEUR-WS.org.

[16] 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. In *MediaEval (CEUR Workshop Proceedings)*, Vol. 2670. CEUR-WS.org.

[17] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, and Julien Morlier. 2020. Classification of Strokes in Table Tennis with a Three Stream Spatio-Temporal CNN for MediaEval 2020. In *MediaEval (CEUR Workshop Proceedings)*, Vol. 2882. CEUR-WS.org.

[18] 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 *ICIP*. IEEE, 554–558.

[19] 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.

[20] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, Akka Zemmari, and Julien Morlier. 2021. *3D Convolutional Networks for Action Recognition: Application to Sport Gesture Recognition.* Springer International Publishing.

[21] Marion Morel, Catherine Achard, Richard Kulpa, and Séverine Dubuisson. 2017. Automatic evaluation of sports motion: A generic computation of spatial and temporal errors. *Image Vis. Comput.* 64 (2017), 67–78.

[22] Hai Nguyen-Truong, San Cao, N. A. Khoa Nguyen, Bang-Dang Pham, Hieu Dao, Minh-Quan Le, Hoang-Phuc Nguyen-Dinh, Hai-Dang Nguyen, and Minh-Triet Tran. 2020. HCMUS at MediaEval 2020: Ensembles of Temporal Deep Neural Networks for Table Tennis Strokes Classification Task. In *MediaEval (CEUR Workshop Proceedings)*, Vol. 2882. CEUR-WS.org.

[23] Soichiro Sato and Masaki Aono. 2020. Leveraging Human Pose Estimation Model for Stroke Classification in Table Tennis. In *MediaEval (CEUR Workshop Proceedings)*, Vol. 2882. CEUR-WS.org.

[24] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding. In *CVPR*. IEEE, 2613–2622.

[25] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. 2020. A Short Note on the Kinetics-700-2020 Human Action Dataset. *CoRR* abs/2010.10864 (2020).

[26] 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).

[27] 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.

[28] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. (2020), 3866–3874.