Sport Task: Fine Grained Action Detection and Classification of Table Tennis Strokes from Videos for MediaEval 2022

Pierre-Etienne Martin[1], Jordan Calandre[2], Boris Mansencal[3], Jenny Benois-Pineau[3], Renaud Péteri[2], Laurent Mascarilla[2] and 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*

Abstract

Sports video analysis is a widespread research topic. Its applications are very diverse, like events detection during a match, video summary, or fine-grained movement analysis of athletes. As part of the MediaEval 2022 benchmarking initiative, this task aims at detecting and classifying subtle movements from sport videos. We focus on recordings of table tennis matches. Conducted since 2019, this task provides a classification challenge from untrimmed videos recorded under natural conditions with known temporal boundaries for each stroke. Since 2021, the task also provides a stroke detection challenge from unannotated, untrimmed videos. This year, the training, validation, and test sets are enhanced to ensure that all strokes are represented in each dataset. The dataset is now similar to the one used in [1, 2]. This research is intended to build tools for coaches and athletes who want to further evaluate their sport performances.

1. Introduction

Action detection and classification is one of the main challenges in computer vision [3]. Throughout the last few years, datasets focusing on action classification have grown tremendously, along with their complexity [2]. There has been a significant number of studies devoted to the analysis of sports gestures using motion-capture systems. Nonetheless, sensors and markers attached to the body have the inherent tendency to interfere with the natural behaviour of the athletes. Therefore, this concern motivates the development of non-invasive methods using video recording from a camera.

The sports video classification project was initiated by the Sports Faculty of the University of Bordeaux (STAPS), the computer science laboratory LaBRI, and the MIA laboratory of the University of La Rochelle [1]. This project intends to develop artificial intelligence and multimedia indexing methods for table tennis stroke recognition. The ultimate goal is to evaluate the performance of individual athletes, especially students, and thereby develop optimal training strategies. For this purpose, we have recorded a video corpus `TTStroke-21` with volunteer athletes.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

mediaeval.sport.task@diff.u-bordeaux.fr (P. Martin)
www.eva.mpg.de/comparative-cultural-psychology/staff/pierre-etienne-martin (P. Martin)
0000-0002-9593-4580 (P. Martin)

[1] This research is supported by the New Aquitania Region's CRISP project - ComputeR vIsion for Sports Performance and the MIRES Federation.

Datasets like UCF-101 [4], HMDB [5, 6], AVA [7] and Kinetics [8, 9, 10] are used in the action recognition field, with an increasing number of video samples and number of classes covered over the years. There are very few datasets available that have a focus on fine-grained classifications in sports, such as FineGym [11] and TTStroke21 [12].

To address the increasing complexity of datasets, some classification methods exploit temporal information as much as possible. For example, [13] learns spatio-temporal dependencies from videos using only RGB data. Alternatively, some methods integrate other modalities extracted from videos, e.g., optical flow [14, 15, 16]. Moreover, in the `TTStroke-21` dataset, stroke classification is challenging, because movements between two strokes share strong visual similarities.

The following sections present the Sport task and its substasks for this year, along with the dataset and the specific terms of use when downloading the dataset. Complementary information on the task may be found on the dedicated page from the MediaEval website[2].

2. Task description

The Sport task is based on the `TTStroke-21` database [17, 12]. This database is a corpus of table tennis recordings with players performing in natural conditions. The dataset delivered through this task focuses on videos acquired with GoPro cameras at a recording speed of 120 frames per second and annotated by professional players. This task offers researchers an opportunity to solve a fine-grained classification problem with videos and annotations of high quality in the sports domain. Compared to the Sport task from MediaEval 2021's edition [18], the dataset has been enriched and the data organization and distribution differ. The task has two subtasks: stroke classification from trimmed videos and stroke detection from untrimmed videos. Each subtask has its own dataset.

Researchers can participate in one or both subtasks and submit up to five runs for each subtask. The participants must fill in the provided XML files dedicated to the test set of the subtask for each run. The content of the XML file varies according to the subtask. The runs have to be submitted in 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 pre-trained networks, was used to compute their runs. The use of pre-trained models on the Sport task dataset `TTStroke-21` of the previous years is however forbidden. The task is considered fully automatic: once the videos are provided to the system, results should be produced without any human intervention. Participants are encouraged to release their code publicly with their submission. This year, similarly to the 2021 edition, a baseline for both subtasks is shared publicly[3] [19].

2.1. Subtask 1 - Stroke Classification

For this subtask, the participants are required to classify a set of trimmed videos containing only one table tennis stroke, or possibly no stroke at all. There are 20 possible stroke classes and an additional non-stroke class. For this purpose, two annotated sets are provided: a training and a validation set with respectively 807 and 230 trimmed videos. A non-annotated test set comprising 118 trimmed videos has to be classified. The trimmed videos in the different sets may have been retrieved from the same untrimmed videos but at different moments in time without overlapping.

[2] https://multimediaeval.github.io/editions/2022/tasks/sportsvideo/

[3] https://github.com/ccp-eva/SportTaskME22

Specifically, the participants are invited to fill an XML file and replace the default label "Unknown" with the stroke class assigned by the participants' method. All submissions will be evaluated in terms of global accuracy for ranking, and detailed with per-class accuracy.

In last year edition, the best global accuracy reached 74.2% [20] using SWIN-Transformers, followed closely by ResNet-50 models (68.8%) beating by a large margin last year baseline (20.4%) [21]. Methods effectiveness seem to be linked to the model architecture, but also the taking into account of both stroke class similarity and class imbalance during training. This year, the task uses the same split of the TTStroke-21 database as in [1, 2] allowing better comparison with previous works outside the MediaEval benchmark scope.

2.2. Subtask 2 - Stroke Detection

For this subtask, the participants are required to segment a set of untrimmed videos with the aim to retrieve strokes whatever the stroke class. For this purpose, two annotated sets are provided: a training set and a validation set, with respectively 16 and 6 untrimmed videos. A non-annotated test set consisting of 6 untrimmed videos has to be temporally segmented. The videos are not shared across the training, the validation, and test sets; however, the same player may appear in the different sets.

Specifically, the participants have to fill the provided test set XML files with the stroke temporal boundaries (frame index of the videos). All submissions will be evaluated in terms of mean Average Precision (mAP) and temporal Intersection over Union (IoU). Both are usually used for image segmentation but are adapted for this task:

- **mAP:** each stroke represents an object to be detected temporally. Detection is considered True when the temporal 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 [22]. This metric will be used for the final ranking of participants.
- **IoU:** the frame-wise overlap between the ground truth and the predicted strokes across all the videos.

For last year's edition, only two participants submitted runs for this difficult subtask [23, 24]. They did not improve the baseline result in terms of mAP but [24] reached an IoU of 0.247 against 0.144 for the baseline, using YOLOv5 model.

3. Dataset description

The dataset was recorded at the Sports Faculty of the University of Bordeaux. It is constituted of player-centred videos without markers or sensors, recorded in natural conditions using GoPro cameras (see Figure 1). Professional table tennis teachers designed a dedicated taxonomy to describe all the possible strokes. The dataset includes 20 table tennis stroke classes: 8 services, 6 offensive strokes, and 6 defensive strokes. The strokes may also be divided in two super-classes: Forehand and Backhand. The dataset was annotated by professional players using a crowd-sourced annotation platform. Non-stroke samples are inferred from the stroke annotations.

In order to be able to share the dataset, we blurred the 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 last year's edition, the classification dataset is enriched this year with new and more diverse video samples. The source videos were trimmed, sorted in class folders and

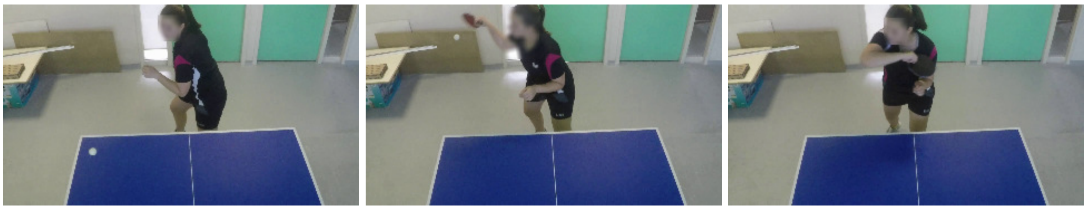

Figure 1: Key frames of a same stroke from TTStroke-21

distributed among train, validation and test sets. A total of 1 155 trimmed videos, representing more than 210 000 frames, are considered for this subtask. For the detection subtask, 100 minutes of table tennis games across 28 videos recorded at 120 frames per second and distributed in train, validation and test sets are considered. It represents more than 718 000 frames. The resolution of the video for both subtasks is 1920 × 1080 representing in total 46.1 GB of disk space. The validation set is provided for each subtask for better comparison across participants. This set may be used for training when submitting the test set's results.

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*.

These *Special Conditions* apply to the use of videos in the scope of the MediaEval workshop task "Sport Task: Fine Grained Action Detection and Classification of Table Tennis Strokes from Videos.". They correspond to the specific usage agreement referred to in the *Usage agreement for the MediaEval 2022 Research Collections*, signed between the user and the University of Delft. The complete acceptance of these *Special Conditions* is a mandatory prerequisite for the provision of the videos as part of the MediaEval 2022 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 January 30th, 2023.

5. Discussions

As last year, the MediaEval Sport task offers two subtasks: i) Classification and ii) Detection of strokes. Classification methods performance has increased since the launch of the task. We hope to have more participation in the detection subtask in order to also improve in the domain of moment of interest in sports. The 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.

Acknowledgment

The authors would like to thank the players, coaches, and annotators who contributed to TTStroke-21.

References

[1] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, 3d attention mechanisms in twin spatio-temporal convolutional neural networks. application to action classification in videos of table tennis games., in: ICPR, IEEE Computer Society, 2021.

[2] P. Martin, 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. thesis, University of La Rochelle, France, 2020. URL: https://tel.archives-ouvertes.fr/tel-03128769.

[3] P. Martin, J. Benois-Pineau, R. Péteri, A. Zemmari, J. Morlier, 3D Convolutional Networks for Action Recognition: Application to Sport Gesture Recognition, Springer International Publishing, 2021.

[4] K. Soomro, A. R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, CoRR abs/1212.0402 (2012).

[5] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: ICCV, IEEE Computer Society, 2011, pp. 2556–2563.

[6] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. J. Black, Towards understanding action recognition, in: ICCV, IEEE Computer Society, 2013, pp. 3192–3199.

[7] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, J. Malik, AVA: A video dataset of spatio-temporally localized atomic visual actions (2018) 6047–6056.

[8] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, CoRR abs/1705.06950 (2017).

[9] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, A. Zisserman, A short note on the kinetics-700-2020 human action dataset, CoRR abs/2010.10864 (2020).

[10] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, A. Zisserman, The ava-kinetics localized human actions video dataset, CoRR abs/2005.00214 (2020).

[11] D. Shao, Y. Zhao, B. Dai, D. Lin, Finegym: A hierarchical video dataset for fine-grained action understanding, in: CVPR, IEEE, 2020, pp. 2613–2622.

[12] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, Fine grained sport action recognition with twin spatio-temporal convolutional neural networks, Multim. Tools Appl. 79 (2020) 20429–20447.

[13] Z. Liu, H. Hu, Spatiotemporal relation networks for video action recognition, IEEE Access 7 (2019) 14969–14976.

[14] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: CVPR, IEEE Computer Society, 2017, pp. 4724–4733.

[15] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 1510–1517.

[16] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, Optimal choice of motion estimation methods for fine-grained action classification with 3d convolutional networks, in: ICIP, IEEE, 2019, pp. 554–558.

[17] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, Sport action recognition with siamese spatio-temporal cnns: Application to table tennis, in: CBMI, IEEE, 2018, pp. 1–6.

[18] P. Martin, J. Calandre, B. Mansencal, J. Benois-Pineau, R. Péteri, L. Mascarilla, J. Morlier, Sports video: Fine-grained action detection and classification of table tennis strokes from videos for mediaeval 2021, in: [25], 2021. URL: http://ceur-ws.org/Vol-3181/paper3.pdf.

[19] P. Martin, Baseline method for the sport task of mediaeval 2022 benchmark with 3d cnns using attention mechanism, in: MediaEval, CEUR Workshop Proceedings, CEUR-WS.org, 2022.

[20] Y. Qian, L. Yu, W. Liu, A. Hauptmann, Learning unbiased transformer for long-tail sports action classification, in: [25], 2021. URL: http://ceur-ws.org/Vol-3181/paper52.pdf.

[21] P. Martin, Spatio-temporal CNN baseline method for the sports video task of mediaeval 2021 benchmark, in: [25], 2021. URL: http://ceur-ws.org/Vol-3181/paper13.pdf.

[22] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: D. J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, volume 8693 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 740–755.

[23] A. Zahra, P. Martin, Two stream network for stroke detection in table tennis, in: [25], 2021. URL: http://ceur-ws.org/Vol-3181/paper55.pdf.

[24] B. J, M. T. T, B. B, J. S, L. N. N, YOLOV5 for stroke detection and classification in table tennis, in: [25], 2021. URL: http://ceur-ws.org/Vol-3181/paper38.pdf.

[25] S. Hicks, K. Pogorelov, A. Lommatzsch, A. G. S. de Herrera, P. Martin, S. Z. Hassan, A. Porter, A. Kasem, S. Andreadis, M. Lux, M. G. Ocaña, A. Liu, M. Larson (Eds.), Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021, volume 3181 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: http://ceur-ws.org/Vol-3181.