Baseline Method for the Sport Task of MediaEval 2023 using 3D CNNs with Attention Mechanisms for Table Tennis Stoke Detection and Classification.

Pierre-Etienne Martin

CCP Department, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany

Abstract

This paper presents the baseline method proposed for the Sports Video task part of the MediaEval 2023 benchmark. This task proposes six sports-related multimedia tasks, each divided into sub-tasks for table tennis and swimming. In this baseline, we focus only on table tennis stroke detection from untrimmed videos (subtask 2.1), and stroke classification from trimmed videos (subtask 3.1). We propose two types of 3D-CNN architectures to solve those two subtasks. Both 3D-CNNs use Spatio-temporal convolutions and attention mechanisms. The architectures and the training process are tailored to solve the addressed subtask. This baseline method is shared publicly online to help the participants in their investigation and alleviate eventually some aspects of the task such as video processing, training method, evaluation, and submission routine. The baseline reaches a mAP of 0.131 and IoU of 0.515 with our v1 model for the detection subtask. For the classification subtask, the baseline method reaches 86.4% of accuracy with our v2 model. The same baseline was used in the 2022 edition. Additional results are incorporated in this paper to encourage comparison and discussion with the participants.

1. Introduction

The field of computer vision has shown considerable interest in the classification of actions from videos [1, 2, 3, 4]. Initially, 2D CNNs were utilized for this task [5, 6], which later evolved into 3D convolution methods to better encapsulate temporal information from videos [7]. The use of optical flow, computed from the RGB stream, was explored to enhance performance and convert RGB variations into movement data [8, 9]. More recently, multi-model methods have been revisited, this time integrating the RGB and audio streams [10], leading to breakthroughs on standard benchmark datasets like Kintetics600 [11].

The MediaEval 2023 Sport Task focuses on the classification and detection of table tennis strokes from videos, as detailed in [12]. The task emphasizes actions with low visual inter-class variability and involves detecting them from untrimmed videos (subtask 2.1) and classifying them from trimmed videos (subtask 3.1). These subtasks are built on the TTStroke-21 dataset [13] and bears similarities to other datasets with low inter-class variability [14, 15, 16, 17].

This baseline is the same as the one presented in the 2022 Mediaeval edition [18, 19**?**] and used in [20]. Its implementation is publicly available on GitHub[1].

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online
pierre_etienne_martin@eva.mpg.de (P. Martin)
www.eva.mpg.de/ccp/staff/pierre-etienne-martin (P. Martin)
0000-0002-9593-4580 (P. Martin)

[1] https://github.com/ccp-eva/SportTaskME23

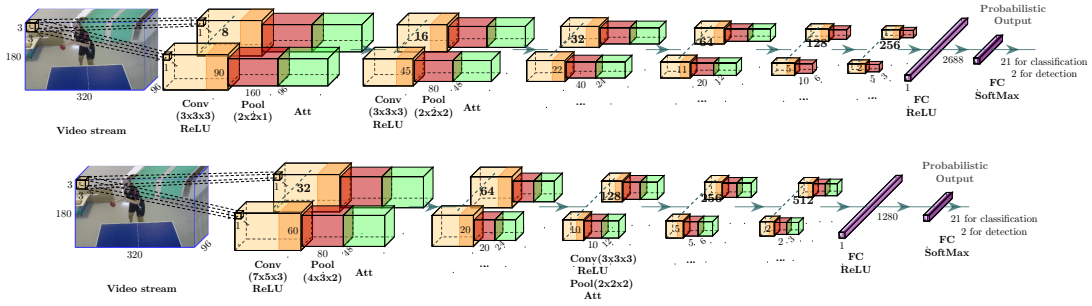

Figure 1: 3D CNNs v1 (top) and v2 (bottom) using Attention Mechanisms for Stroke Classification and Detection.

2. Method

The proposed method leverages solely the RGB data from the given videos, drawing inspiration from [21]. A key difference is the lack of Region Of Interest (ROI) computation from Optical Flow values. The RGB frames are resized to a width of 320 and stacked to form 96-length tensors, either from the trimmed videos or according to the annotation boundaries. Data augmentation techniques, such as starting at different time points and spatial transformations (flip and rotation), are employed to enhance variability.

Two versions of the method, V1 and V2 as illustrated in Figure 1, are utilized. V1 consists of a sequence of four conv+pool+attention layers followed by two conv+pool layers. All convolutional layers employ 3x3x3 filters. The initial layers use 2x2x1 pooling filters, while the subsequent layers use 2x2x2 pooling filters. V2, on the other hand, comprises a sequence of five conv+pool+attention layers. The convolution filters for the first two blocks are 7x5x3 in size, with 4x3x2 pooling filters. The remaining blocks use 3x3x3 and 2x2x2 filters for convolution and pooling, respectively.

The training process employs Nesterov momentum over a set number of epochs, with the learning rate adjusted based on loss evolution [21]. The model that performs best on the validation loss is retained. The same training methods are used for both subtasks. The objective function is the cross-entropy loss of the softmax-processed output, summed over the batch:

$$\mathcal{L}(y, class) = -log(\frac{exp(y'_{class})}{\sum_i^N exp(y_i)}) \tag{1}$$

For classification, we consider 21 classes, and for detection, two classes, as in [22]. Negative samples are used for detection, and testing involves trimmed proposals or a sliding window across the entire video. Strokes shorter than 30 frames are ignored. The classification-trained model is also tested on detection without additional training. Two approaches are considered: comparing the negative class score against all others, and comparing the negative class score against the sum of all others. Various decision methods are tested. For more details, see [23].

3. Results

This section outlines the results for each subtask based on the metrics detailed in [12]. Both subtasks involved training the models for 2000 epochs with a learning rate of .0001, a momentum of .5, and a weight decay of .005.

3.1. Subtask 2.1 - Table Tennis Stroke Detection

Table 1's first section presents results using video candidates from the test set, which are non-overlapping, successive 150-frame samples from the test videos. The primary evaluation metric is mAP, with the V2 model using a Vote decision performing best. However, this method of extracting video candidates is not efficient for stroke detection. For better segmentation, a sliding window with a step of one is used on the test videos, and outputs are combined using the previously mentioned window methods. Models from subtask 1 (marked with †) are also tested. The second part of Table 1 shows some improvement, with the V1 model achieving the best mAP and IoU scores using the segmentation method, but the V2 models do not perform as well.

Table 1
Models performance on detection subtask in terms of mAP | IoU with proposals (first hald) and with sliding window segmentation on the test set (second half.

Model	No Window	Vote	Mean	Gaussian
V1	.111 | .358	.114 | .360	.113 | **.365**	.113 | .361
V2	.111 | .322	**.118** | .329	.117 | .333	.117 | .331
V1	-	**.131** | **.515**	.00201 | .341	.00227 | .33
V1†	-	.00012 | .201	.0.0017 | .210	.00428 | .211
V1††	-	.000019 | .203	.000054 | .203	.00174 | .205
V2	-	.000731 | .308	.102 | .473	.1 | .466
V2†	-	.000506 | .207	.00173 | .215	.00237 | .216
V2††	-	.00145 | .209	.00185 | .211	.00261 | .212

† Negative class VS all †† Negative class VS sum of all

3.2. Subtask 3.1 - Table Tennis Stroke Classification

As reported in Table 2, V1 and V2 perform similarly on the stroke classification subtask, but V2 using the Gaussian window decision performs the best with 86.4% of accuracy on the test set. This model finished convergence at epoch 815 with train and validation accuracies of .989 and .813 respectively. The confusion matrices of this run are depicted in Figure 2.

Table 2
Models performance on classification subtask in terms of accuracy

Model	No Window	Vote	Mean	Gaussian
V1	.847	.839	.856	.856
V2	.856	.822	.831	**.864**

As we can notice on the confusion matrix, the model has the tendency to classify some strokes as non-strokes (negative class). This is certainly due to the variation in the negative class, increasing its dedicated latent space and giving more probability to the unseen samples to fall in it. This could be solved by increasing the variability of these samples via data augmentation or more recording of these strokes.

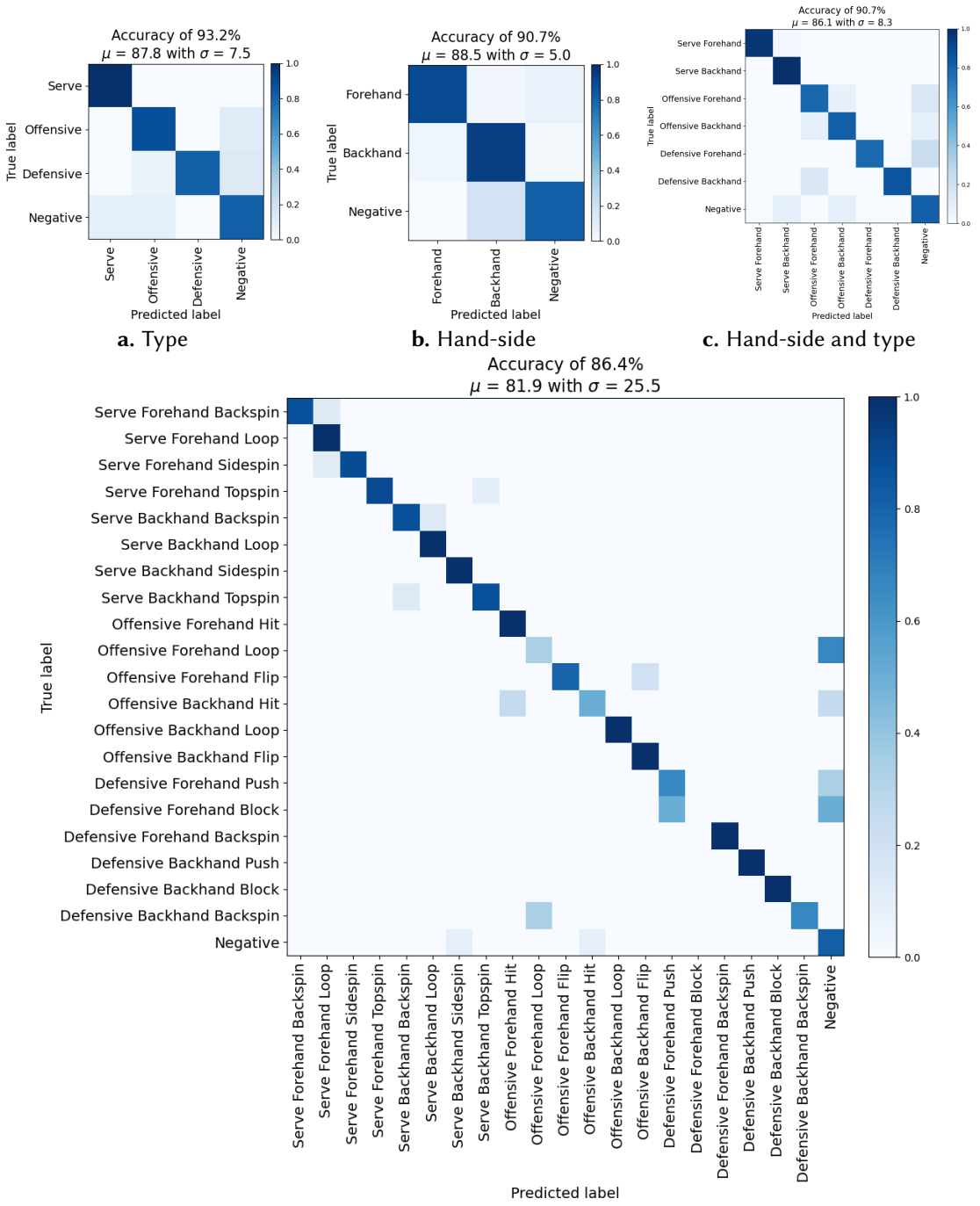

Figure 2: Confusion matrices of the best classification run on the test set with different granularities.

4. Conclusion

This baseline aims to assist participants in the Sports Video Task, building on last year's baseline [**?**]. This paper provides more results than the previous year to foster discussion and comparison. Enhancements can be made by integrating insights from subtasks 2.1 and 3.1, refining the training process with more complex data augmentation or weighted loss. For the next edition, we plan to provide a baseline for the entire sport task, including table tennis and swimming.

References

[1] 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).

[2] 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.

[3] 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).

[4] A. J. Piergiovanni, M. S. Ryoo, Avid dataset: Anonymized videos from diverse countries, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020.

[5] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, Action recognition with dynamic image networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 2799–2813.

[6] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: NIPS, 2014, pp. 568–576.

[7] T. Lima, B. J. T. Fernandes, P. V. A. Barros, Human action recognition with 3d convolutional neural network, in: LA-CCI, IEEE, 2017, pp. 1–6.

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

[9] 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.

[10] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, Y. Choi, Merlot reserve: Multimodal neural script knowledge through vision and language and sound, in: CVPR, 2022.

[11] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, CoRR abs/1808.01340 (2018).

[12] A. Erades, P. Martin, R. V. B. Mansencal, R. Péteri, J. Morlier, S. Duffner, J. Benois-Pineau, SportsVideo: A multimedia dataset for event and position detection in table tennis and swimming, in: Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online, 1-2 February 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2023.

[13] 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.

[14] 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.

[15] Y. Li, Y. Li, N. Vasconcelos, RESOUND: towards action recognition without representation bias, in: ECCV (6), volume 11210 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 520–535.

[16] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, M. Wray, Scaling egocentric vision: The EPIC-KITCHENS dataset, CoRR abs/1804.02748 (2018).

[17] S. Noiumkar, S. Tirakoat, Use of optical motion capture in sports science: A case study of golf swing, in: ICICM, 2013, pp. 310–313.

[18] S. Hicks, A. G. S. de Herrera, J. Langguth, A. Lommatzsch, S. Andreadis, M. Dao, P. Martin, A. Hürriyetoglu, V. Thambawita, T. S. Nordmo, R. Vuillemot, M. A. Larson (Eds.), Working Notes Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023, volume 3583 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3583.

[19] P. Martin, J. Calandre, B. Mansencal, J. Benois-Pineau, R. Péteri, L. Mascarilla, J. Morlier, Sport task: Fine grained action detection and classification of table tennis strokes from videos for mediaeval 2022, in: [18], 2022. URL: https://ceur-ws.org/Vol-3583/paper26.pdf.

[20] L. Hacker, F. Bartels, P. Martin, Fine-grained action detection with RGB and pose information using two stream convolutional networks, in: [18], 2022. URL: https://ceur-ws.org/Vol-3583/paper21.pdf.

[21] 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.

[22] P. Martin, Spatio-temporal cnn baseline method for the sports video task of mediaeval 2021 benchmark, in: MediaEval, CEUR Workshop Proceedings, CEUR-WS.org, 2021.

[23] P. Martin, Fine-Grained Action Detection and Classification from Videos with Spatio-Temporal Convolutional Neural Networks. Application to Table Tennis., Ph.D. thesis, University of La Rochelle, France, 2020. URL: https://tel.archives-ouvertes.fr/tel-03128769.