

Baseline Method for the Sport Task of MediaEval 2022 with 3D CNNs using Attention Mechanisms

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 2022 benchmark. This task proposes two subtasks: stroke classification from trimmed videos, and stroke detection from untrimmed videos. This baseline addresses both subtasks. We propose two types of 3D-CNN architectures to solve the 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 method reaches 86.4% of accuracy with our v2 model for the classification subtask. For the detection subtask, the baseline reaches a mAP of 0.131 and IoU of 0.515 with our v1 model.

1. Introduction

Action classification from videos is a popular topic in the computer vision field [1, 2, 3, 4]. In order to solve such task, 2D CNNs were first introduced [5, 6]. Then, to better capture the temporal information from videos, 3D convolution methods emerged [7, 8]. Optical flow computed from the RGB stream was also investigated in order to boost performance and translate RGB changes into movement information [9, 10]. Recently, multi-model methods are re-investigated but this time combining the RGB and the audio streams [11] leading to the state-of-the-art on common benchmark datasets such as Kintetics600 [12]. Multi-view methods combined with Transformers [13] are also the current state-of-the-art in many action classification dataset [14, 15].

In the Sport Task of MediaEval 2022, the focus is on the classification and detection of table tennis strokes from videos. As described in [16], the task focuses on low visual inter-class variability actions: classify them from trimmed videos (subtask 1) and detect them from untrimmed videos (subtask 2). The task is based on TTStroke-21 dataset [17] and is similar to other datasets with low inter-class variability [18, 19, 14, 20].

This baseline, publicly available on GitHub¹, tackles the two subtasks and aims to help participants in their submission such as the processing of the videos, the annotation files and the deep learning methods.

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

✉ pierre_etienne_martin@eva.mpg.de (P. Martin)

🌐 www.eva.mpg.de/comparative-cultural-psychology/staff/pierre-etienne-martin (P. Martin)

🆔 0000-0002-9593-4580 (P. Martin)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://github.com/ccp-eva/SportTaskME22>

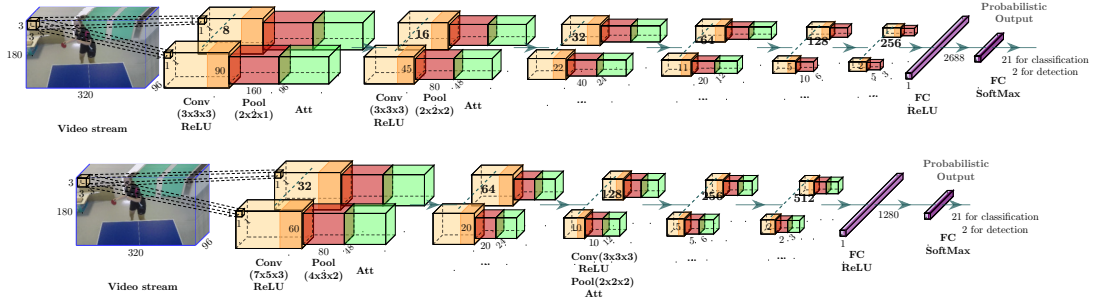


Figure 1: 3D CNNs v1 (top) and v2 (bottom) using Attention Mechanisms for Stroke Classification and Detection.

2. Method

The method has been kept simple and uses only the RGB information from the provided videos. The implementation is inspired from [21]. The main divergence is the absence of Region Of Interest (ROI) which was computed from Optical Flow values. The data processing is trivial: The RGB frames are resized to a width of 320 and stacked together to form tensors of length 96 either from the trimmed videos or following the annotation boundaries available in the XML files. Data are augmented to increase variability: start at different time points and spatial transformations (flip and rotation).

Two versions, V1 and V2 are introduced and depicted in Figure 1. V1 is a sequence of four conv+pool+attention layers and two conv+pool layers. All convolutional layers use 3x3x3 filters. The first layers use 2x2x1 pooling filters (no pooling on the temporal domain) and 2x2x2 pooling filters for the other layers. V2 is a sequence of five conv+pool+attention layers. Conv. filters are of size 7x5x3 and pooling filters of size 4x3x2 for the first two blocks. The remaining blocks use 3x3x3 and 2x2x2 for conv. and pooling filters respectively. V2 leads to almost squared feature maps after the second block so that horizontality, verticality and temporality can be better combined before the fully connected layers.

The training method uses Nesterov momentum over a fixed amount of epochs. The learning rate is modified according to the loss evolution [21]. The model with the best performance on the validation loss is saved. The training methods are the same for both subtasks. The objective function is the cross-entropy loss of the output processed by the softmax function summing over the batch:

$$\mathcal{L}(y, class) = -\log\left(\frac{\exp(y'_{class})}{\sum_i \exp(y_i)}\right) \quad (1)$$

We consider 21 classes for the classification task and two classes for the detection task as previously done in [22]. Negative samples are extracted for the detection task and negative proposals are built on its test set. Testing is performed with the trimmed proposal (with one window centered or with a sliding window and several post-processing approaches) or by running a sliding window on the whole video for the detection task. The latest output is processed in order to segment framewise the strokes. Too short strokes, less than 30 frames, are not considered. The model trained on the classification task is also tested on the detection task without further training on the detection data. Two approaches are considered: 1) Negative class score VS all others for decision and 2) Negative class score VS sum of all the others. Several decision methods are also tested: No Window, Vote, Mean, and Gaussian according to a temporal window. See [23] for further details.

Table 1

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

Table 2

Models performance on detection subtask in terms of mAP | IoU with proposals on the test set

Model	No Window	Vote	Mean	Gaussian
V1	.111 .358	.114 .360	.113 .365	.113 .361
V2	.111 .322	.118 .329	.117 .333	.117 .331

3. Results

This section presents the results per subtask according to the metrics presented in [16]. For the two subtasks, we trained the models for 2000 epochs using a learning rate of .0001, a momentum of .5 and a weight decay of .005.

3.1. Subtask 1 - Stroke Classification

As reported in Table 1, 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 matrix of this run is depicted in Figure 2.

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.

3.2. Subtask 2 - Stroke Detection

Table 2 reports the results using video candidates from the test set. Video candidates are simply non-overlapped successive samples of length 150 frames from the test videos. The main metric for evaluation is the mAP, and therefore the model V2 using a Vote decision performs the best. However, extracting video candidates in such way is not efficient to detect the strokes. That is why in Table 3 results using another segmentation methods are reported.

To perform a better segmentation, a sliding window with step one is used on the test videos. The outputs are combined in order to make a decision following the same previously presented window methods. The models from subtask 1 are also tested.

As we can see, the segmentation method allows the model V1 to reach the best performance in terms of mAP and IoU. However it is not the case for the V2 models.

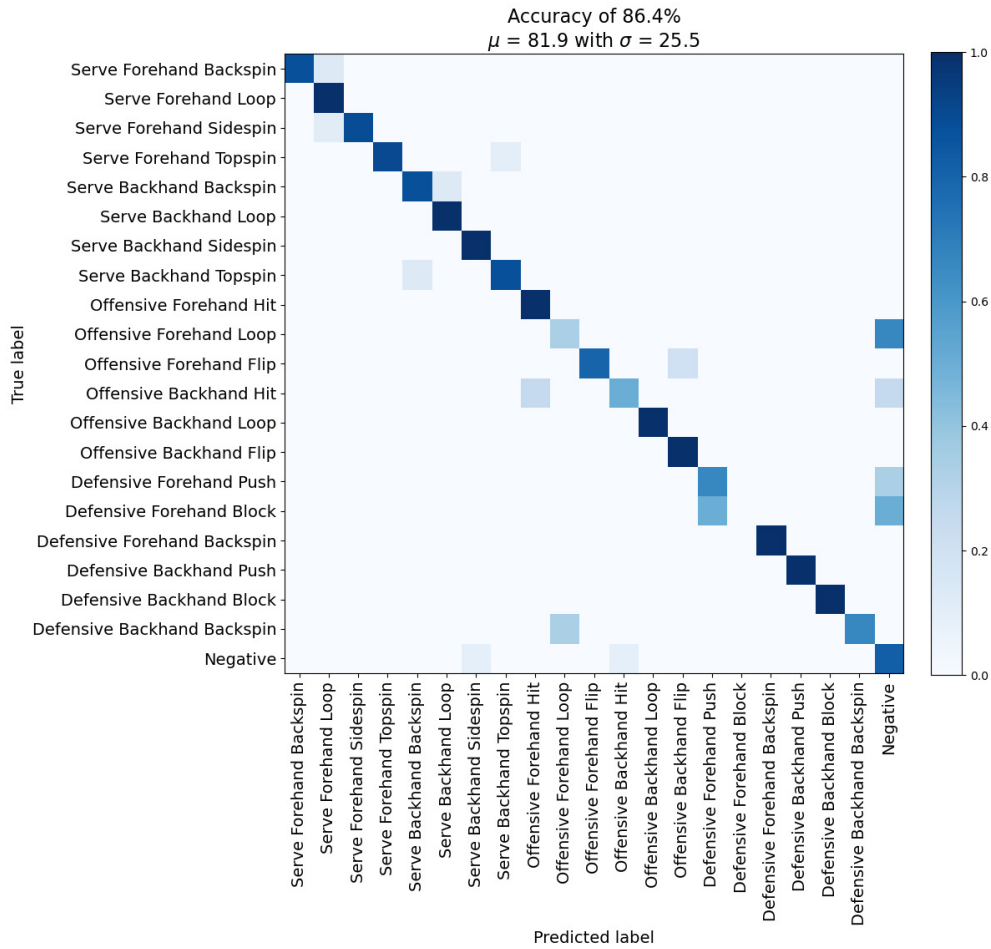


Figure 2: Confusion matrix of the best classification run on the test set.

Table 3

Models performance on detection subtask in terms of mAP | IoU with sliding window segmentation

Model	Vote	Mean	Gaussian
V1	.131 .515	.00201 .341	.00227 .33
V2	.000731 .308	.102 .473	.1 .466

4. Conclusion

This baseline intends to help the participants solving the Sports Video Task. This work is in the continuity of last year’s baseline [22] and more tools were implemented to help the participants as suggested at the last edition. Improvements can be made by combining knowledge from subtask 1 to solve subtask 2. Also, the data augmentation and the loss can be improved to balance the unbalanced distribution of the samples. Finally, the segmentation method for stroke detection can still be improved to boost the performance in this subtask. These possibilities of improvements may be implemented in next year’s baseline.

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, virtual, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/c28e5b0c9841b5ef396f9f519bf6c217-Abstract.html>.
- [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] H. J. Kim, J. S. Lee, H. S. Yang, Human action recognition using a modified convolutional neural network, in: ISNN (2), volume 4492 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 715–723.
- [8] 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.
- [9] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: CVPR, IEEE Computer Society, 2017, pp. 4724–4733.
- [10] 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.
- [11] 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.
- [12] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600, CoRR abs/1808.01340 (2018).
- [13] M. Dehghani, A. Gritsenko, A. Arnab, M. Minderer, Y. Tay, Scenic: A jax library for computer vision research and beyond, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21393–21398.
- [14] 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).
- [15] M. Monfort, C. Vondrick, A. Oliva, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. M. Brown, Q. Fan, D. Gutfreund, Moments in time dataset: One million videos for event understanding, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 502–508.
- [16] 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: MediaEval, CEUR Workshop Proceedings, CEUR-WS.org, 2022.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] S. Noiumkar, S. Tirakoat, Use of optical motion capture in sports science: A case study of golf swing, in: ICICM, 2013, pp. 310–313.
- [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. (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>.