Spatio-Temporal CNN Baseline Method for the Sports Video Task of MediaEval 2021 Benchmark

Pierre-Etienne Martin
CCP Department, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany
pierre_etienne_martin@eva.mpg.de

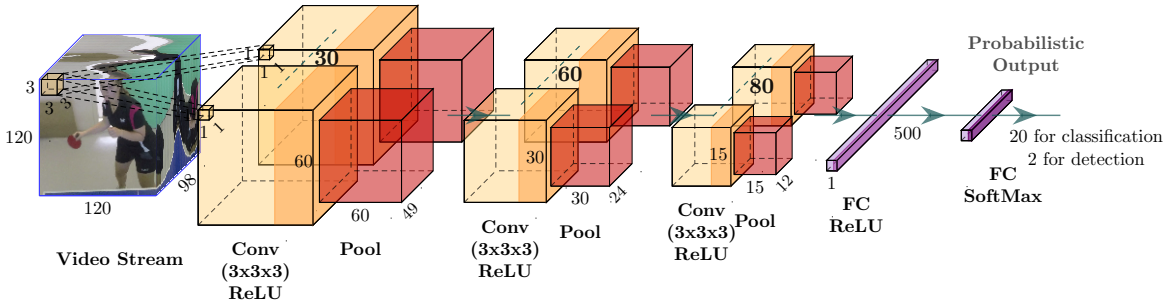


Figure 1: Spatio-Temporal CNN architecture for Stroke Classification and Detection.

ABSTRACT

This paper presents the baseline method proposed for the Sports Video task part of the MediaEval 2021 benchmark. This task proposes a stroke detection and a stroke classification subtasks. This baseline addresses both subtasks. The spatio-temporal CNN architecture and the training process of the model are tailored according to the addressed subtask. The method has the purpose of helping the participants to solve the task and is not meant to reach state-of-the-art performance. Still, for the detection task, the baseline is performing better than the other participants, which stresses the difficulty of such a task.

1 INTRODUCTION

Most recent action detection and classification methods developed in the literature have been using deep learning approaches and high-dimensional spaces. In the domain of image classification, a specific kind of Neural Network has become very popular: the Convolutional Neural Networks (CNNs). Since the breakthrough at the 2012 ImageNet Challenge, CNNs have demonstrated a great improvement for image classification.

For video applications in general and action recognition in particular, the first models proposed were a direct extension of image classification methods [1, 20] using 2D convolutions. However, to better capture the temporal information proper to video content, the use of 3D convolutions has emerged [7, 8]. One can also consider temporal information using the motion extracted from successive frames, such as the optical flow. The latest can be used i) as a single modality or in parallel with the RGB information [2, 3, 20, 21]; or ii) to train a network for extracting motion features to perform classification at a later stage [4]. These methods also raise the question of how to fuse the different modalities [5, 10]. In [9, 19], the estimated pose is used jointly with these two modalities to perform action classification. In [15] all the three modalities are used and fused in order to perform stroke classification.

MediaEval'21, December 13-15 2021, Online

As part of the task organization, for the first time since the beginning of the Sports Video task (in 2019 [11]), we decided to provide a baseline to alleviate minor aspects of the task, such as video and xml processing; and help the participants in their submission. The baseline method uses a 3D CNN inspired from [13, 14]. We adjusted the method to answer both proposed subtasks of this year's edition [12]: stroke detection and stroke classification from videos of the `TTStroke-21` corpus. The implementation of the method is available publicly on Github[1].

2 METHOD

In order to perform classification and detection, we consider the model architecture presented in Fig. 1. For each subtask, a distinct model has been trained on the train set. We train both using a stochastic gradient approach with a Nesterov momentum of 0.5 [16], a weight decay of 0.005 [6] and a constant learning rate of 0.0001. Both models are trained over 500 epochs. The objective function is the cross-entropy loss of the output processed by the softmax function (eq. 1) summing over the batch:

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

At each epoch, the model is validated on the validation set. The model performing the best on this set is saved and then evaluated on the test set. The model is fed with the video frames resized to 120×120 and staked successively in cuboids of length 98, representing approximately 0.82 seconds.

For the detection task, we inferred *Non-stroke* segments from the annotated *Stroke* segments. We considered only segments between two consecutive strokes greater than 200 frames. Such a segment is divided in successive blocks of 200 frames, non overlapping, and added has a negative sample for training the model. The split using 200 frames allows a correct number of negative samples: from the 783 train and 234 validation segments, we inferred respectively for each set 1196 and 260 negative segments. No negative segments have been inferred from the test set. Stroke detection is tackled

[1] https://github.com/ccp-eva/SportTaskME21

as a classification task by considering two classes: *Stroke* on *Non-stroke*. From the test set, which has no temporal boundaries, we created window proposals of length 150 every 150 frames for all the videos. This size was chosen empirically and meant to be revised to achieve good performance. For the classification task, all the classes were not represented in the dataset but we still consider all the 20 possible stroke classes.

To train the model, we inputted the RGB cuboids composed of the successive frames from the *starting frame* of the considered segment. The desired output is the class vector summing to one and binary at training time. Its length is the number of considered classes: 2 for detection and 20 for classification. Each element represents the probability of belonging to a class. During inference, we follow a similar procedure, and the class decision is the *argmax* of the output vector.

3 RESULTS

This section presents the results per subtask according to the metrics presented in [12].

3.1 Subtask 1 - Stroke Detection

The detection subtask was tackled as a classification task, considering the strokes and non-strokes samples. After 500 epochs, the model reached 98.3% and 75.7% of accuracy, respectively, on the train and validation sets. On the test set, the model is evaluated using the mAP metric. This metric takes into account the number of actions detected and their overlapping with the ground truth. The baseline achieves an mAP of 0.0173, which the two participants of this subtask did not outperform.

Runs are also evaluated using a global IoU that considers only the frame-wise overlap of the detected strokes with the ground truth annotations. The number of strokes detected is no longer taken into account in the evaluation. The baseline achieves a Global IoU of 0.144, which was outperformed by one participant.

The method's performance is quite low due to the method being relatively simple. It also relies on a straightforward and non-efficient window proposal to segment the strokes without fusing the output decision. Indeed, two consecutive windows, part of the same stroke and classified as strokes, will be classified as two different strokes and not a single one, and will therefore have an impact on the mAP metric. The method can easily be improved by considering better proposals and fusing the output decisions.

3.2 Subtask 2 - Stroke Classification

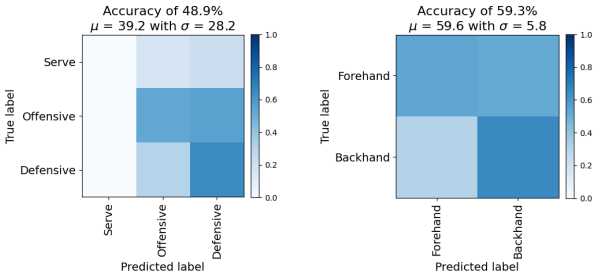
The results for the stroke classification subtask on the test set are reported in the table 1. This table is divided into different sections for considering different refined classifications. After training, the model reached only 25.2% and 28.9% of accuracy, respectively, on the train and validation sets.

- "Global" consider all the 20 classes
- "Type" consider only the type of the stroke: Defensive, Offensive or Service
- "Hand-Side" consider only Forehand and Backhand superclasses
- "Type and Hand-Sided" consider the intersection of the two last clusters leading to 6 classes.

The confusion matrix of the "Type" and "Hand-Side" are also depicted in Fig. 2 for further analysis.

Table 1: Baseline performance in term of accuracy (%).

Global	Type and Hand-Sided	Type	Hand-Side
20.4	33	48.9	59.3



a. Type **b.** Hand-side

Figure 2: Confusion matrices with higher level categories.

From table 1, we can state that the performance of the baseline, considering all classes, is limited. This may be improved by further analysis of the corpus and further training. Indeed, only 18 classes over the 20 possible were present in the corpus this year, which simplifies the complexity of the task and could have been taken into account in the model's design. Fig. 2.a reveals that the services have not been learned at all, which is undoubtedly due to the input processing during training which considers only the 100 first frames and is therefore unable to capture features from these longer strokes. Finally, Fig. 2.b underlines the main weakness of the model: being unable of distinguishing Forehand and Backhand strokes. The pipeline's method could consider higher level categories, following a cascade method, to improve the performance. Two of the three participants have outperformed by far the baseline performance [17, 18].

4 CONCLUSION

This baseline intends to help the participants solving the Sports Video Task. The baseline performance remains limited, but its publicly available implementation allows the participants to not start from scratch. Many aspects of the method may be improved, such as the data processing: a spatial and temporal ROI may increase the performance. Similarly with the architecture of the model, which was kept very simple, or the training method that could have merged the train and validation sets before inferring on the test set.

The detection subtask seems to be challenging. No participants were able to beat the baseline performance with regard to the mAP metric, which is the ranking metric. This subtask is new in the Sports Video Task, which also explains the low results obtained. However we believe much improvement can be obtained since our method has tackled it as a classification task. The window proposal is also very crude and can easily be improved.

The classification subtask has gathered more participants with, overall, more successful performance. This may be explained by the task's non-novelty in the history of the MediaEval benchmark and the more active investigation in this field.

Next year we plan to gather ideas from this year's submissions to improve the baseline and give a more substantial base to the new participants joining the Sports Video Task.

REFERENCES

[1] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. 2018. Action Recognition with Dynamic Image Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 12 (2018), 2799–2813.

[2] Jordan Calandre, Renaud Péteri, and Laurent Mascarilla. 2019. Optical Flow Singularities for Sports Video Annotation: Detection of Strokes in Table Tennis. In *MediaEval (CEUR Workshop Proceedings)*, Vol. 2670. CEUR-WS.org.

[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] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. 2019. MARS: Motion-Augmented RGB Stream for Action Recognition. In *CVPR*. IEEE Computer Society, 7882–7891.

[5] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *CVPR*. IEEE Computer Society, 1933–1941.

[6] Stephen Jose Hanson and Lorien Y. Pratt. 1988. Comparing Biases for Minimal Network Construction with Back-Propagation. In *NIPS*. 177–185.

[7] Ho Joon Kim, Joseph S. Lee, and Hyun Seung Yang. 2007. Human Action Recognition Using a Modified Convolutional Neural Network. In *ISNN (2) (Lecture Notes in Computer Science)*, Vol. 4492. Springer, 715–723.

[8] Tiago Lima, Bruno J. T. Fernandes, and Pablo V. A. Barros. 2017. Human action recognition with 3D convolutional neural network. In *LA-CCI*. IEEE, 1–6.

[9] Diogo C. Luvizon, David Picard, and Hedi Tabia. 2018. 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning. In *CVPR*. IEEE Computer Society, 5137–5146.

[10] 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

[11] 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.

[12] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Laurent Mascarilla, Jordan Calandre, and Julien Morlier. 2021. Sports Video: Fine-Grained Action Detection and Classification of Table Tennis Strokes from videos for MediaEval 2021. In *MediaEval (CEUR Workshop Proceedings)*. CEUR-WS.org.

[13] 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.

[14] 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*. IEEE, 1–6.

[15] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2021. Three-Stream 3D/1D CNN for Fine-Grained Action Classification and Segmentation in Table Tennis. *CoRR* abs/2109.14306 (2021). arXiv:2109.14306 https://arxiv.org/abs/2109.14306

[16] Yurii E. Nesterov. 2004. *Introductory Lectures on Convex Optimization - A Basic Course*. Applied Optimization, Vol. 87. Springer.

[17] Trong-Tung Nguyen, Thanh-Son Nguyen, Gia-Bao Dinh Ho, Hai-Dang Nguyen, and Minh-Triet Tran. 2021. HCMUS at MediaEval 2021: Ensembles of Action Recognition Networks with Prior Knowledge for Table Tennis Strokes Classification Task. In *MediaEval (CEUR Workshop Proceedings)*. CEUR-WS.org.

[18] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G. Hauptmann. 2021. Learning Unbiased Transformer for Long-Tail Sports Action Classification. In *MediaEval (CEUR Workshop Proceedings)*. CEUR-WS.org.

[19] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2020. LCR-Net++: Multi-Person 2D and 3D Pose Detection in Natural Images. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 5 (2020), 1146–1161.

[20] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS*. 568–576.

[21] Xuanhan Wang, Lianli Gao, Peng Wang, Xiaoshuai Sun, and Xianglong Liu. 2018. Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length. *IEEE Trans. Multimedia* 20, 3 (2018), 634–644.