Two Stream Network for Stroke Detection in Table Tennis

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

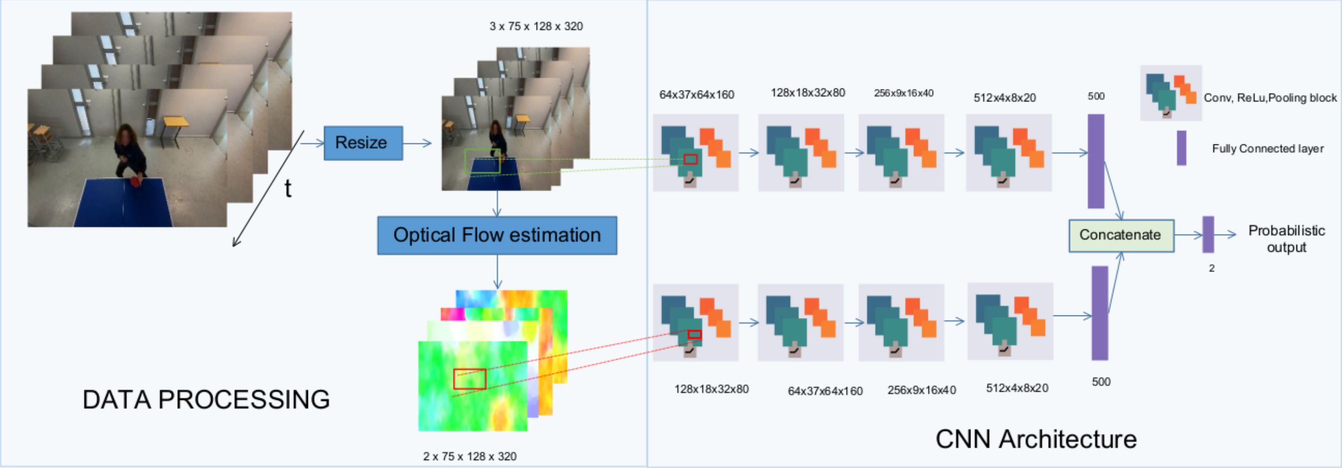

Figure 1: Pipeline method for stroke detection from videos. Cuboids of RGB and optical flow are fed to the network and classified as stroke or non-stroke. The feature dimension is described as follow: $RGBchannels \times temporal \times height \times width$**.**

ABSTRACT

This paper presents a table tennis stroke detection method from videos. The method relies on a two-stream Convolutional Neural Network processing in parallel the RGB Stream and its computed optical flow. The method has been developed as part of the MediaEval 2021 benchmark for the Sport task. Our contribution did not outperform the provided baseline on the test set but has performed the best among the other participants with regard to the mAP metric.

1 INTRODUCTION

With the advent of Convolutional Neural Networks (CNNs), especially after the success of AlexNet [9], object detection, localization, and classification from images and videos have greatly progressed [3, 5, 9, 22].The development of computer vision methods has motivated broader applications in the academic world. Our team is currently working on egocentric recordings from children in kindergarten and at home. The analysis of these recordings shall give us an automatic overview of their interactions on a daily basis. We hope to link these interactions with their cognitive development and, thereby, better understand early child development. With our participation in the Sports Video Task [12], in the stroke detection subtask, we hope to perfect our knowledge in event detection and transpose it to our project.

The diversity of applications and visual data in sport, makes sports video analysis attractive for researchers. Automated sport event detection and action classification, especially from low-resolution videos, are helpful for monitoring and training purposes. For example in [6, 10], the authors automate the performance analysis for the training optimization of players. Similarly, Sports Video task at MediaEval 2021 benchmark aims at improving athlete performance and training experience through the first steps of stroke detection and classification from videos.

Event detection in videos is the first step to many other hot-topics such as video summarizing [7], automated semantic segmentation [1] and action recognition [4, 16]. These methods may be used to build summary, selecting highlights, and assisting players in training sessions. One way to approach the problem of event detection in sports with balls, can be through ball detection and tracking. Several researchers have tried to get the 2D, and 3D ball trajectories in order to achieve so [17, 18, 20].

Inspired from [8, 14, 15, 21], this method combines the optical flow and features learned from the RGB stream in order to detect a stroke in table tennis and assess its duration. This implementation is an extension of the baseline code provided by the Sport Task organizers [11].

2 APPROACH

Initially, we sought to use ball detection and tracking to perform stroke detection. The first implementation used the pretrained model TTNet [21]. However, the model failed to adapt to the acquisition conditions from TTStroke-21 [13], on which the task is built upon, and no fine-tuning was possible since no ball coordinates are available in the provided annotations. Therefore we decided to train a model from scratch.

MediaEval'21, December 13-15 2021, Online

In this section, we first present the preparation of the videos and then the model presenting the processed data. Both processes are depicted in Fig. 1. Post processing is performed to form a final decision.

2.1 Data Preparation

In video content analysis, the motion of objects of interest between frames can be of significant interest in order to understand their evolution in space. As such, we decided to use optical flow as a modality to perform stroke detection. Inspired by [14], we decided to use DeepFlow method [23] to compute the optical flow from consecutive frames. The optical flow is computed from frames resized to 320 × 128. This size was initially chosen to keep the ball at least two pixels big, as it has previously been done in [21]. Both the RGB and optical flow frames are consecutively stacked in a tensor of length 75. As in [11], stroke detection is tackled as a classification problem with two classes: "Stroke" and "Non-stroke".

2.2 Model

As shown in figure 1, our Two-Stream model is composed of two branches of the same length. Each branch is a succession of four blocks and each block is composed of a convolutional layer with $3 \times 3 \times 3$ filters, followed by a ReLU activation function, and a $2 \times 2 \times 2$ pooling layer. The output of each branch is then flattened and fed to a fully connected layer that outputs a feature vector of length 500. Both feature vectors are then concatenated and fed into a final fully connected layer of length two to predict the "Stroke" and "Non-storke" classes. One branch takes RGB frames of the video and the other computed optical flow. The model is trained using a stochastic gradient descent method over 250 epochs with a learning rate of 0.001, a batch size of 10, a weight decay of 0.005, and a Nesterov momentum [19] of 0.5. The negative samples creation and input processing is the same as the baseline [11].

2.3 Post Processing

Our model classifies 75 consecutive frames. In order to create stroke segments over the whole video, we classify every 75 frames of the videos, which leads to applying a sliding window without overlap. If two consecutive segments are classified as stroke, the segments are fused to create only one stroke.

3 RESULTS AND ANALYSIS

The metrics for evaluating the detection performance are described in [12]. Our approach reached a mean Average Precision (mAP) of 0.00124 and a Global Intersection over Union (G-IoU) of 0.0700. It falls behind the baseline which reaches respectively 0.0173 and 0.144. Our other attempts using early concatenation of the RGB and Optical Flow modalities - meaning an input of size $5 \times 320 \times 128$ in one branch model - or training method without shuffling of the data, reached even lesser performance.

Nevertheless, from a classification point of view, and according to the Fig. 2, our model learned the stroke features and can perform reasonable results when stroke boundaries are known: 86.4% of accuracy on the validation set after only 60 epochs. Which may indicates that the main failure is coming from the post processing method.

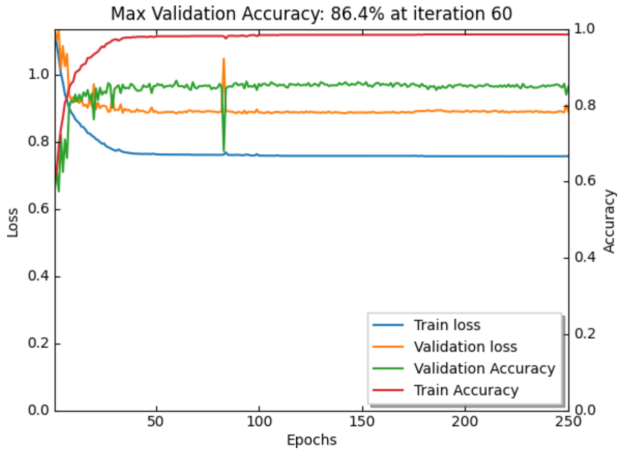

Figure 2: Training Process

Table 1: Stroke concentration and duration in frame per set.

Set	# Strokes/1K frames	Mean	Min	Max
Train	1.85	143.2 ± 36.16	52	296
Valid	2.28	134.3 ± 26.13	72	292
Test	0.57	361.0 ± 770.7	75	4500

Indeed, by looking at the stroke distribution across the different sets, see table 1, we may notice how little the inferred stroke ratio is on the test set: 0.57 strokes for 1000 frames, whereas the stroke rate is 1.85 and 2.28 for 1000 frames in the training and validation sets. Furthermore, our post processing was not limited in term of stroke duration, leading to everlasting strokes: 4500 frames - meaning the fusions of 60 consecutive video segments. These points indicate that our post-processing method can be improved.

A better separation of the stroke may be reached by defining the event using ball tracking and the ball motion [2]. This was our initial attempt, inspired by [21], but the available pretrained model considers a different point of view and was unable to adapt to the `TTStroke-21` videos point of view.

4 CONCLUSION

The Sports Video Task, and more specifically the stroke detection subtask, has proven to be challenging. Even if our implementation has learned to classify strokes, we were not able to outperform the baseline performance. We have underlined the importance of the post processing step through a stroke concentration and duration analysis. Furthermore, our failure to adapt a pretrained model on similar dataset, but with a different acquisition point of view, stresses the difficulty of the deep trained models to adapt to a change of scene, which is inherent to the fine-grained aspect of the classification subtask. As first time participants, we thought to tackle only one task to ease our submission. However, we now believe that a method tackling both the detection and classification may be the best for solving the Sport Video subtasks.

REFERENCES

[1] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. 2011. Event detection and recognition for semantic annotation of video. *Multimedia tools and applications* 51, 1 (2011), 279–302.

[2] Jordan Calandre, Renaud Péteri, Laurent Mascarilla, and Benoit Tremblais. 2021. Table Tennis ball kinematic parameters estimation from non-intrusive single-view videos. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–6.

[3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[4] Chandni J Dhamsania and Tushar V Ratanpara. 2016. A survey on human action recognition from videos. In *2016 online international conference on green engineering and technologies (IC-GET)*. IEEE, 1–5.

[5] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[6] Mike D Hughes and Roger M Bartlett. 2002. The use of performance indicators in performance analysis. *Journal of sports sciences* 20, 10 (2002), 739–754.

[7] Yasmin S Khan and Soudamini Pawar. 2015. Video summarization: survey on event detection and summarization in soccer videos. *International Journal of Advanced Computer Science and Applications* 6, 11 (2015), 256–259.

[8] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, and others. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[10] Adrian Lees. 2003. Science and the major racket sports: a review. *Journal of sports sciences* 21, 9 (2003), 707–732.

[11] 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.

[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, Renaud Péteri, and Julien Morlier. 2018. Sport Action Recognition with Siamese Spatio-Temporal CNNs: Application to Table Tennis. In *CBMI*. IEEE, 1–6.

[14] 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 *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*. IEEE, 554–558. https://doi.org/10.1109/ICIP.2019.8803780

[15] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. 3D attention mechanisms in Twin Spatio-Temporal Convolutional Neural Networks. Application to action classification in videos of table tennis games.. In *25th International Conference on Pattern Recognition (ICPR2020) - MiCo Milano Congress Center, Italy, 10-15 January 2021*.

[16] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. Fine grained sport action recognition with twin spatio-temporal convolutional neural networks. *Multimedia Tools and Applications* 79, 27 (2020), 20429–20447.

[17] Hnin Myint, Patrick Wong, Laurence Dooley, and Adrian Hopgood. 2015. Tracking a table tennis ball for umpiring purposes. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 170–173.

[18] Hnin Myint, Patrick Wong, Laurence Dooley, and Adrian Hopgood. 2016. Tracking a table tennis ball for umpiring purposes using a multi-agent system. (2016).

[19] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*. PMLR, 1139–1147.

[20] Sho Tamaki and Hideo Saito. 2013. Reconstruction of 3d trajectories for performance analysis in table tennis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1019–1026.

[21] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 884–885.

[22] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*. 3551–3558.

[23] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. 2013. DeepFlow: Large Displacement Optical Flow with Deep Matching. In *2013 IEEE International Conference on Computer Vision*. 1385–1392. https://doi.org/10.1109/ICCV.2013.175