Four-stream network and Dynamic Images for Sports Video Classification: Classification of Strokes in Table Tennis

Jordan Calandre[1], Renaud Péteri [2], Laurent Mascarilla[3]
[1]MIA Laboratory, La Rochelle University, France
{jordan.calandre1,renaud.peteri,lmascari}@univ-lr.fr

ABSTRACT

In this working note, results for the MediaEval 2020 Sports Video Annotation "Detection of Strokes in Table Tennis" task are presented. Fine-grained action classification remains a complex task due to the low variance between two strokes, especially in natural conditions. Our proposal is therefore based on motion, which is the most obvious representation of what players are doing. Motion information is captured at the image level by optical flow streams and summarized at the sequence level by Dynamic Images that encode temporal information. A multiple stream architecture is presented, combining RGB-based Dynamic Images, Dynamic Images based on optical flow, and RGB frames to classify table tennis strokes.

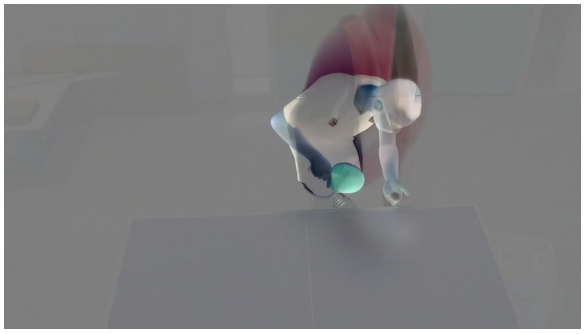

Figure 1: Dynamic Image

1 INTRODUCTION

Fine-grained action recognition in natural conditions remains difficult even after the success of CNN architectures for image and video processing. Datasets like UCF-101 [11], or HMDB [7] are useful for benchmarking methods classifying human action into a given set of sport classes, however the fine-grained recognition of gestures of a specific sport leads to new challenges.

The dataset `TTStroke-21` [10] is made up for this purpose and is much more challenging than most previous datasets. Acquisition is done using standard cameras, without depth maps or motion capture information. The number of strokes are also heavily unbalanced, which can lead to overfitting when training deep neural networks.

Deep learning methods for 2D images recognition tasks led to the spread of CNN network for video analysis. Popular methods, like 3D-CNN, using 3D filters instead of 2D filters on video frames, require huge datasets to be trained efficiently. An alternative method is to use the optical flow. These approaches like two-stream networks or Siamese Networks have been very successful. The optical flow represents the movement between two consecutive frames, but without estimating long term dependencies. The movement being the obvious representation of a stroke, we focus on this feature to enhance our previous proposal [3]. Optical flow and Dynamic Images [1] are used to capture image motion information.

2 OUR APPROACH

We have participated to MediaEval 2019 [9] with a method using optical flow singularities [3], and have noticed that temporal data were not fully exploited with this approach. Our new proposal for MediaEval 2020 [8] is to use Dynamic Images [1] to summarize each sequence based on RGB, along with optical flow obtained by the so-called PWC-Network [12]. A multiple-stream architecture with late fusion is then used to process the different network inputs.

2.1 Dynamic Images (DI)

A Dynamic Image [1] (DI) is a representation of an image sequence in a single frame. This frame is obtained by representing the video using a ranking function on its frames [5]. A pixel pooling operation is applied with the ranking function to average the pixel values over time.

2.2 Dynamic Optical Flow (DOF)

The optical flow being a two dimensional vector field that represents the apparent motion between two consecutive frames, it does not capture long-term motion. When combining the flow of each frame of a video sequence using the same approach as for DI, the motion of an entire stroke into a single image is aggregated, and thus long-term interactions can be captured.

To obtain a dense flow with clean boundaries, the PWC-Network [12] has been selected as it achieves suitable results at decent speed. It has been trained using the Sintel dataset [2].

Since the videos at hand contain compression artifacts, a Gaussian filter is applied before estimating the optical flow.

2.3 CNN Architecture

The proposed CNN architecture is composed of up to four branches. Each branch corresponds to a ResNet[6] with 152 layers, pretrained on ImageNet [4] but the input type varies according to the branch. The five possible inputs for the branches are: A Dynamic Image (DI) computed on the whole sequence; the RGB frame from the middle of the sequence; two Dynamic Images computed on each half of the sequence (DIHalf); a Dynamic Image computed on the optical flow (DOF). The input type, of the branches, for each run is presented in Table. 1. Every input is a `224x224` pixel image, cropped around the

MediaEval'20, December 14-15 2020, Online

Figure 2: Dynamic Optical Flow (DOF)

Table 1: Run results

Method	Train set	Val Set	Test set
DI	25.00%	25.70%	11.58%
DI + RGB	30.34%	23.65%	10.17%
2*DIHalf	62.28%	36.48%	11.58%
2*DIHalf + RGB	63.05%	36.48%	11.51%
2*DIHalf + DOF + RGB	79.21%	44.58%	12.99%

player using `Detectron2` [13]. We modified the last fully connected layer to have 20 neurons, which is the number of considered classes. To combine the branches outputs, a late fusion is applied followed by a fully connected layer that results in the final stroke classification score.

The network was trained over 100 epochs, with a learning rate of 0.05 and a momentum of 0.9 using 10-folds cross validation. All the video sequences of the dataset with at least two different strokes are used in the validation set.

3 RESULTS AND ANALYSIS

The accuracy, for each of the five allowed runs of the task, is presented in Table. 1 for training validation and and testing sets.

To our surprise, the scores are quite similar for runs using one DI or two DIHalf. By averaging the features only on half the sequence, the use of two DIHalf (runs 3,4 and 5) was expected to better represent the movement. This seems to have no real impact on the overall result, nor the adding of the RGB frame located at the middle of the sequence. The only run with a better score is the one with DOF (Dynamic Optical Flow). The DOF encodes the movement but unlike the dynamic RGB images, it provides an insight of the direction of the players hands/grip when in action.

In last year task challenge, using optical flow singularities [3], our best score was 50/354 by adding a weight on the predicted strokes to compensate the unbalanced dataset. We obtained 46/354 correctly classified moves for the two best runs without class-weighted SVM.

Compared to last year, our network has a better estimate of the drive's type (Forehand vs Backhand) presented in Table. 2. We also considerably increased player's stroke estimate (Serve/Offensive/Defensive). This metric increased from 48.87% to 65.25%. The confusion matrix for the drive and stroke estimation, for run 5, is presented in Fig.3.

Table 2: Comparison of the accuracy between our MediaEval 2019 and MediaEval 2020 submissions

Metric	2019	2020
Drive(Forehand/Backhand)	61.58%	65.25%
Group(Serve/Offensive/Defensive)	48.87%	65.82%
Group and Drive	29.10%	49.15%
Total accuracy	14.12%	12.99%

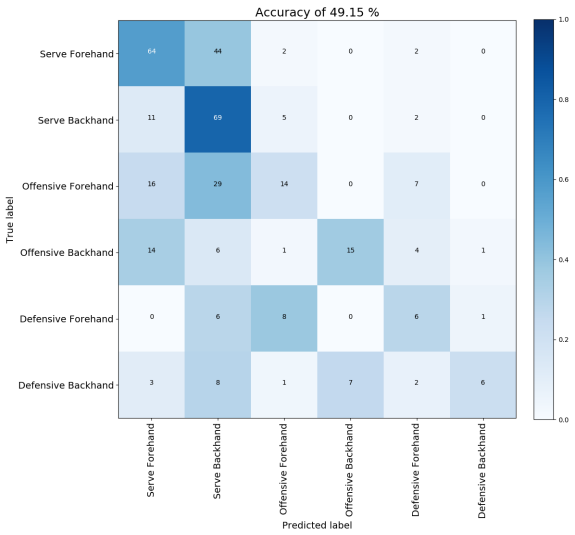

Figure 3: Accuracy of the predicted drive and stroke estimates

4 DISCUSSION AND OUTLOOK

This paper presents the approach of the MIA laboratory for the Sports Video Annotation on single-sport dataset task. Due to the difficulty of the task, such as rare classes samples and different camera viewpoints, the overfit obtained during the training sessions leads to a low score, but it gives an insight of what kind of information is missing in the proposed Dynamic Images. RGB frames and Dynamic Images are arbitrarily split in the middle of each sequence, but an impact detection of the ball could be used to make a more meaningful splitting. Lastly, unbalanced data must be better handled as prediction is clearly biased toward some stroke classes.

5 ACKNOWLEDGMENTS

The research is supported by the Region of Nouvelle Aquitaine through the CRISP project and by the CNRS MIRES federation.

REFERENCES

[1] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. 2016. Dynamic Image Networks for Action Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016-December. 3034–3042.

https://doi.org/10.1109/CVPR.2016.331

[2] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. 2012. A naturalistic open source movie for optical flow evaluation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-33783-3_44

[3] Jordan Calandre, Renaud Péteri, and Laurent Mascarilla. 2019. Optical flow singularities for sports video annotation: Detection of strokes in table tennis. In *CEUR Workshop Proceedings*, Vol. 2670.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[5] Basura Fernando and Stephen Gould. 2017. Discriminatively Learned Hierarchical Rank Pooling Networks. *International Journal of Computer Vision* (2017). https://doi.org/10.1007/s11263-017-1030-x arXiv:1705.10420

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2016.90 arXiv:1512.03385

[7] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre Thomas. 2013. Hmdb51: A large video database for human motion recognition. In *High Performance Computing in Science and Engineering 12: Transactions of the High Performance Computing Center, Stuttgart (HLRS) 2012*. IEEE Computer Society, 571–582. https://doi.org/10.1007/978-3-642-33374-3

[8] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Laurent Mascarilla, Jordan Calandre, and Julien Morlier. 2020. Sports Video Classification: Classification of Strokes in Table Tennis for MediaEval 2020. In *Proc. of the MediaEval 2020 Workshop, Online, 14-15 December 2020*.

[9] 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 *CEUR Workshop Proceedings*.

[10] Pierre Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks: Application to table tennis. *Multimedia Tools and Applications* (2020). https://doi.org/10.1007/s11042-020-08917-3

[11] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR* abs/1212.0402 (2012). arXiv:1212.0402 http://arxiv.org/abs/1212.0402

[12] Deqing Sun, Xiaodong Yang, Ming Yu Liu, and Jan Kautz. 2018. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* abs/1709.0 (2018), 8934–8943. https://doi.org/10.1109/CVPR.2018.00931 arXiv:1709.02371

[13] Uxin Wu, Alexander Kirillov, Francisco Massa, Wan-YenLo, and Ross Girshick. 2019. Detectron2. *https://github.com/facebookresearch/detectron2* (2019).