Classification of Strokes in Table Tennis with a Three Stream Spatio-Temporal CNN for MediaEval 2020

Pierre-Etienne Martin[1], Jenny Benois-Pineau[1], Boris Mansencal[1], Renaud Péteri[2], Julien Morlier[3]
[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, Talence, France
[2]MIA, La Rochelle University, La Rochelle, France
[3]IMS, University of Bordeaux, Talence, France
mediaeval.sport.task@diff.u-bordeaux.fr

ABSTRACT

This work presents a method for classifying table tennis strokes using spatio-temporal convolutional neural networks. The fine-grained classification is performed on trimmed video segments recorded at 120 fps with different players performing in natural conditions. From those segments, the frames are extracted, their optical flow is computed and the pose of the player is estimated. From the optical flow amplitude, a region of interest is inferred. A three stream spatio-temporal convolutional neural network using combination of those modalities and 3D attention mechanisms is presented in order to perform classification.

1 INTRODUCTION

Recognition of actions with low inter-class variability remains a challenge [2, 8, 16, 18]. The target application of our research is fine-grained action recognition in sports with the aim of improving athletes performance [3, 9, 21]. The purpose is to make cameras "smart" to analyse sport practices [1, 4, 19]. The first step here is to classify strokes played in incoming video streams.

Based on our previous works [14], we propose a method using RGB and optical flow data to perform classification[1]. Without loss of generality, we are interested in recognition of strokes in table tennis through the MediaEval 2020 Sport task [11], based on TTStroke-21 dataset [14]. Compared to our work at MediaEval 2019 [12] for the same task [10], our method differs by the use of the estimated pose and attention mechanism [15] based on [5, 20]. The difficulty of this task is to find characteristics for each class of strokes using a limited dataset. In this paper, we present in section 2 a three stream network aiming at extracting features with enough inter-class discrimination to perform classification. Section 3 presents the results and conclusion is drawn in section 4.

2 PROPOSED APPROACH

To deal with the low inter-class variability of TTStroke-21, the most complete information from video must be used, i.e. both appearance (RGB) and motion (Optical Flow). Spatio-temporal convolutions were performed on cuboids of RGB frames and on cuboids of Optical Flow (OF). Those two kinds of information were processed simultaneously through a Twin architecture [15]. A third branch with temporal convolutions was added to handle the estimated pose. The extracted frames from videos of size (1920×1080) were resized to (320×180).

2.1 Optical Flow estimation

As presented in [13], flow estimators and its normalization can strongly impact classification. We used Dense Inversive Search estimator [6] because of its computational speed. Each OF frame $\mathbf{V} = (v_x, v_y)$ was encoded with horizontal v_x and vertical v_y motion computed from two consecutive RGB frames. The estimated OF was smoothed with a Gaussian filter with kernel size 3×3 and then multiplied by the computed foreground [22] to keep only foreground motion.

2.2 Estimation of the Region of interest

The region of interest (ROI) center $\mathbf{X_{roi}} = (x_{roi}, y_{roi})$ was estimated from the maximum of the OF $\mathbf{V}$ norm and the center of gravity of all pixels with non-null OF norm as follows:

$$\begin{aligned}
\mathbf{X_{max}} &= (x_{max}, y_{max}) = \underset{x,y}{argmax}(||\mathbf{V}||_1) \\
\mathbf{X_g} &= (x_g, y_g) = \frac{1}{\sum\limits_{\mathbf{X}\in\Omega} \delta(\mathbf{X})} \sum_{\mathbf{X}\in\Omega} \mathbf{X}\delta(\mathbf{X}) \\
\text{with } \delta(\mathbf{X}) &= \begin{cases} 1 & \text{if } ||\mathbf{V}(\mathbf{X})||_1 \neq 0 \\ 0 & \text{otherwise} \end{cases} \\
x_{roi} &= \alpha\, f_{\omega_x}(x_{max},\ W) + (1-\alpha)\, f_{\omega_x}(x_g,\ W) \\
y_{roi} &= \alpha\, f_{\omega_y}(y_{max},\ H) + (1-\alpha)\, f_{\omega_y}(x_g,\ H)
\end{aligned} \tag{1}$$

with parameter $\alpha = 0.6$, set empirically, $\Omega = (\omega_x, \omega_y) = (320, 180)$ the size of video frames. Function $f_\omega(u, S) = max(min(u, \omega - \frac{S}{2}),\ \frac{S}{2})$ allows to have data inputted to our network within the region of interest. To avoid jittering within our RGB and OF cuboids, of size $(W \times H \times T) = (120 \times 120 \times 98)$, a Gaussian filter with kernel size k_{size} and with scale parameter $\sigma_{blur} = 0.3 * ((k_{size} - 1) * 0.5 - 1) + 0.8$ was applied along the temporal dimension to average the center position. In our experiments, the optimal kernel size was found to be $\frac{1}{3}$ second which represents $k_{size} = 41$ frames at 120 fps.

2.3 Pose estimation

The pose was computed from single RGB images using the PoseNet model [17]. Its implementation is available online[2]. It supplies poses and human joints positions and their score. We discard some human joints that are not visible in the considered videos such as the knees and the ankles. The 13 human joints considered are thus

[1] This work was supported by the New Aquitania Region through CRISP project - ComputeR vIsion for Sport Performance and the MIRES federation.

MediaEval'20, 14-15 December 2020, Online

[2] https://github.com/rwightman/posenet-python

the nose, both eyes, ears, shoulders, elbows, wrists and hips. The pose coordinates (mean of the joint coordinates) and its score are also taken into account leading to a descriptor vector of length $N_{joints} = 14$. Even if the faces are blurred, its joints are still well located. Other players may appear in the scene background, which lead to the detection of several poses in the same frame. In this case, the closest pose, from center of the previously computed ROI, was considered. If no pose is detected, the descriptor vector is filled with ROI center coordinates and a score of 0.

2.4 Data normalization

The RGB data were normalized to map their value into the interval $[0, 1]$. Following [13], the OF was normalized using the mean μ and standard deviation σ of the maximum absolute values distribution of each OF components over the whole dataset as described in equation 2:

$$\begin{aligned} v' &= \frac{v}{\mu + 3\times\sigma} \\ v^N(i,j) &= \begin{cases} v'(i,j) & \text{if } |v'(i,j)| < 1 \\ SIGN(v'(i,j)) & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

with v and v^N representing respectively one component of the OF $\mathbf{V}$ and its normalization. This normalization method maps the values into interval [-1,1] and increases the magnitude of most vectors making the OF more relevant for classification.

2.5 Model architecture

The model was similar to the Twin Spatio-Temporal Convolutional Neural Network - TSTCNN with attention mechanisms presented in [15]. It comprises two branches with three 3D convolutional layers with 30, 60, 80 filters respectively, followed by a fully connected layer of size 500. They take respectively cuboids of RGB values and OF of size $(W \times H \times T)$. The 3D convolutional layers use $3 \times 3 \times 3$ space-time filters with a dense stride and padding of 1 in each direction. Their output is processed by max-pooling layers using kernels of size $2 \times 2 \times 2$. Each max-pooling layer feeds an attention block. An extra branch processing the pose data of size $(N_{joints} \times T) = (14 \times 98)$ is added. It follows the same organization than the two other branches, but without attention mechanism and uses 1D convolutions and max-pooling along the temporal dimension. The three branches are fused two by two using bilinear fully connected layers ($y = x_1^T A x_2 + b$) of size 20, which represent the number of classes. The three resultant outputs are summed and processed by a Softmax function to output probabilistic scores used for classification.

2.6 Data augmentation

Data augmentation was made online, generating different inputs at each epoch during training phase. Each stroke sample was fed to the model once per epoch. For temporal augmentation, T successive data from the RGB, OF and Pose modalities, were extracted following a normal distribution around the center of the stroke video segment with standard deviation of $\sigma = \frac{\Delta t - T}{6}$. Spatial augmentation was performed with random rotation in the range ±10°, random translation in range ±0.1 in x and y directions, random homothety in range 1 ± 0.1 and flip in horizontal direction with 0.5 of probability. The OF and Pose values were updated accordingly. Transformations were applied on the region of interest avoiding crops outside the image borders. During the test phase, no augmentation was performed and the T extracted frames were temporally centered on the stroke segment.

2.7 Training phase

All models were trained from scratch. Due to early overfitting, only 200 epochs were used for training the models using all the training samples. The optimization method was a stochastic gradient descent with Nesterov momentum of 0.5, with learning rate of 0.001, weight decay of 0.05 and a batch size of 5. The objective function was the cross-entropy loss.

3 RESULTS

Five runs on the test set were submitted. Run 1 corresponds to the decision from the proposed model with temporally centered features on the stroke, so called "Coarse". Run 2, 3 and 4 correspond to the same model but using a temporal sliding window on the stroke segments for decision making. The runs correspond respectively to the "Vote" rule, "Avg" rule and "Gaussian" rule. The reader can refer to [14] for further details. Run 5 corresponds to decision of the RGB-branch with attention mechanism. The bilinear layer becomes then a simple linear layer.

Table 1: Runs performances in term of accuracy (%).

Runs	1	**2**	3	4	5
Models	**Three stream STCNN**				RGB-STCNN
Decision	Coarse	**Vote**	Avg	Gaussian	Coarse
Accuracy	24.3	**26.6**	25.4	25.4	20.3

In general, classification results are very low compared to the ones obtained in [15]. This is due to the lower amount of videos for this task and the different split of the dataset: for this task, strokes for the train and test sets are extracted from different videos, which is not the case in [15].

From Table 1, best performances are obtained using all modalities with vote rule decision. This underlines the importance of modality fusion within the architecture and the gain of considering the whole stroke, and not only the $T = 98$ centered frames. Moreover, by merging stroke classes such as the drive: "Forehand", "Backhand"; the context: "Serve", "Offensive", "Defensive"; or their combination (6 classes); run 2 obtains respectively 72.3%, 76.8% and 60.7% of accuracy. The higher scores prove the capacity of the model to learn the characteristics of Table Tennis games. Surprisingly, the context is better classified than the drive.

4 CONCLUSION

Our submission is ranked 2^{nd} in the Sport Task of MediaEval 2020 [11]. The obtained results are better than last year, with a slightly modified dataset. The use of Pose information and attention mechanism allowed such improvements. However, the global accuracies remain low certainly because of the limited amount of samples used for training our models. The challenging task of fine-grained action recognition from few video samples remains open.

REFERENCES

[1] Amin Ahmadi, Edmond Mitchell, Chris Richter, François Destelle, Marc Gowing, Noel E. O'Connor, and Kieran Moran. 2015. Toward Automatic Activity Classification and Movement Assessment During a Sports Training Session. *IEEE Internet Things J.* 2, 1 (2015), 23–32. https://doi.org/10.1109/JIOT.2014.2377238

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. *CoRR* abs/1804.02748 (2018). arXiv:1804.02748 http://arxiv.org/abs/1804.02748

[3] Christopher J. Ebner and Rainhard Dieter Findling. 2019. Tennis Stroke Classification: Comparing Wrist and Racket as IMU Sensor Position. In *MoMM 2019: The 17th International Conference on Advances in Mobile Computing & Multimedia, Munich, Germany, December 2-4, 2019*, Pari Delir Haghighi, Ivan Luiz Salvadori, Matthias Steinbauer, Ismail Khalil, and Gabriele Anderst-Kotsis (Eds.). ACM, 74–83. https://doi.org/10.1145/3365921.3365929

[4] Moritz Einfalt, Dan Zecha, and Rainer Lienhart. 2018. Activity-Conditioned Continuous Human Pose Estimation for Performance Analysis of Athletes Using the Example of Swimming. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. IEEE Computer Society, 446–455. https://doi.org/10.1109/WACV.2018.00055

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

[6] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. 2016. Fast Optical Flow Using Dense Inverse Search. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9908. Springer, 471–488. https://doi.org/10.1007/978-3-319-46493-0_29

[7] Martha A. Larson, Steven Alexander Hicks, Mihai Gabriel Constantin, Benjamin Bischke, Alastair Porter, Peijian Zhao, Mathias Lux, Laura Cabrera Quiros, Jordan Calandre, and Gareth Jones (Eds.). 2020. *Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019*. CEUR Workshop Proceedings, Vol. 2670. CEUR-WS.org. http://ceur-ws.org/Vol-2670

[8] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. RESOUND: Towards Action Recognition Without Representation Bias. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11210. Springer, 520–535. https://doi.org/10.1007/978-3-030-01231-1_32

[9] Ruichen Liu, Zhelong Wang, Xin Shi, Hongyu Zhao, Sen Qiu, Jie Li, and Ning Yang. 2019. Table Tennis Stroke Recognition Based on Body Sensor Network. In *Internet and Distributed Computing Systems - 12th International Conference, IDCS 2019, Naples, Italy, October 10-12, 2019, Proceedings (Lecture Notes in Computer Science)*, Raffaele Montella, Angelo Ciaramella, Giancarlo Fortino, Antonio Guerrieri, and Antonio Liotta (Eds.), Vol. 11874. Springer, 1–10. https://doi.org/10.1007/978-3-030-34914-1_1

[10] 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, See [7]. http://ceur-ws.org/Vol-2670/MediaEval_19_paper_6.pdf

[11] 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*.

[12] 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, See [7]. http://ceur-ws.org/Vol-2670/MediaEval_19_paper_58.pdf

[13] 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

[14] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks. *Multim. Tools Appl.* 79, 27-28 (2020), 20429–20447. https://doi.org/10.1007/s11042-020-08917-3

[15] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2021. 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*. IEEE Computer Society.

[16] S. Noiumkar and S. Tirakoat. 2013. Use of Optical Motion Capture in Sports Science: A Case Study of Golf Swing. In *ICICM*. 310–313.

[17] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. 2018. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11218. Springer, 282–299. https://doi.org/10.1007/978-3-030-01264-9_17

[18] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding. (2020), 2613–2622. https://doi.org/10.1109/CVPR42600.2020.00269

[19] Wan-Lun Tsai. 2018. Personal Basketball Coach: Tactic Training through Wireless Virtual Reality. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR 2018, Yokohama, Japan, June 11-14, 2018*, Kiyoharu Aizawa, Michael S. Lew, and Shin'ichi Satoh (Eds.). ACM, 481–484. https://doi.org/10.1145/3206025.3206084

[20] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. In *CVPR*. IEEE Computer Society, 6450–6458.

[21] Kun Xia, Hanyu Wang, Menghan Xu, Zheng Li, Sheng He, and Yusong Tang. 2020. Racquet Sports Recognition Using a Hybrid Clustering Model Learned from Integrated Wearable Sensor. *Sensors* 20, 6 (2020), 1638. https://doi.org/10.3390/s20061638

[22] Zoran Zivkovic and Ferdinand van der Heijden. 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognit. Lett.* 27, 7 (2006), 773–780. https://doi.org/10.1016/j.patrec.2005.11.005