Three-Stream 3D/1D CNN for Fine-Grained Action Classification and Segmentation in Table Tennis

Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, Julien Morlier

Three-Stream 3D/1D CNN for Fine-Grained Action Classification and Segmentation in Table Tennis.

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

Jenny Benois-Pineau
Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800
F-33400, Talence, France
jenny.benois-pineau@u-bordeaux.fr

Renaud Péteri
MIA, La Rochelle University
La Rochelle, France
renaud.peteri@univ-lr.fr

Julien Morlier
IMS, University of Bordeaux
Talence, France

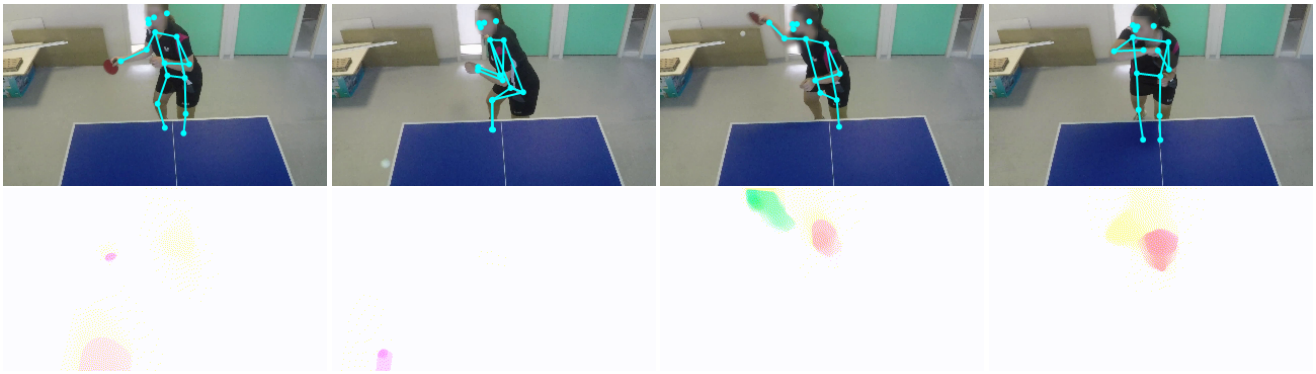


Figure 1: Frames of an "Offensive Forehand Hit" stroke from `TTStroke-21` with its estimated pose and optical flow.

ABSTRACT

This paper proposes a fusion method of modalities extracted from video through a three-stream network with spatio-temporal and temporal convolutions for fine-grained action classification in sport. It is applied to `TTStroke-21` dataset which consists of untrimmed videos of table tennis games. The goal is to detect and classify table tennis strokes in the videos, the first step of a bigger scheme aiming at giving feedback to the players for improving their performance. The three modalities are raw RGB data, the computed optical flow and the estimated pose of the player. The network consists of three branches with attention blocks. Features are fused at the latest stage of the network using bilinear layers. Compared to previous approaches, the use of three modalities allows faster convergence and better performances on both tasks: classification of strokes with known temporal boundaries and joint segmentation and classification. The pose is also further investigated in order to offer richer feedback to the athletes.

MMSports '21, October 20, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8670-8/21/10.
https://doi.org/10.1145/3475722.3482793

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Activity recognition and understanding**; 3D imaging; Computer vision; **Computer vision problems**.

KEYWORDS

Action Classification, Spatio-temporal Convolutions, Table Tennis, Movement analysis, Multi-modal fusion

ACM Reference Format:
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.. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports (MMSports '21), October 20, 2021, Virtual Event, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3475722.3482793

1 INTRODUCTION AND RELATED WORKS

Fine-grained action classification is being more and more investigated in recent years due to its various potential applications such daily living care [3, 5], video security and surveillance [27] or in sport activities [10, 17, 25]. The difference with coarse-grained action classification [9, 28, 30] lays in the high intra-class similarity of the actions. Movements performed are often similar since they focus on one particular activity. Moreover, since videos are recorded in the same context, the background scene and manipulated objects are similar in all videos. Consequently, all possible information

should be extracted from the performed movement itself in order to discriminate actions. The target application of our research is fine-grained action recognition in sports with the aim of improving athletes performance.

Collecting individual data from athletes by body-worn sensors (connected watches, smart clothes, exoskeleton) might be a valuable source of information for classification of similar actions. However, the analysis of gestures is often confined to laboratory studies [6, 13, 34]. Sound has also proven to be efficient for event detection [1] but may not be used for more complex tasks. In [31], the authors propose an advanced real-time solution for scene segmentation, ball trajectory estimation and event detection but are not considering stroke classification.

The use of pose, expressed as coordinates of skeleton joints, has also become popular for action recognition. In [32], the authors apply 3D CNNs on gesture recognition with RGB-D data. They use only four frames mixing 3D and 2D convolutions and max pooling. Joint information is fused using a Deep Belief Network [19]. Similarly, PoTion [4] uses movement of the human joints as features to improve the classification score of the I3D models [2]. Pose has also been used in sport: [8] performs classification of four classes of football (soccer) actions based on pose estimation. The authors of [26] investigate shot direction in Tennis using the pose of the player. In [15], the authors propose a multi-task method for 2D/3D pose estimation and action recognition. Similarly, [24], based on LRC-Net [23], builds a pseudo ground truth for 3D poses from images using 2D pose search in a projected 3D pose dataset in order to offer 3D human pose from images. In [35], pose representations from the pose estimators are feed to a 3D CNN in order to obtain spatio-temporal representation used for action classification. Furthermore, a 3D attention mechanism has been investigated on joint skeleton using LSTM [12]. The pose can also be used for spatial segmentation as in [29].

However, there are limitations of using skeleton-based approaches for action recognition as pointed out in [37]. The authors manage to induce large errors with attacks on the pose based models through low variation of the inputted pose. The use of several modalities is therefore needed to be less dependent on the pose estimation alone. Pose can also be used for further analysis: in [33] the position of a table tennis ball is predicted according to the player's pose. In [20], qualitative measures of tennis and karate gestures are computed for comparing the pose of expert and novice participants.

Recording of "markerless" and "sensorless" video of performing athletes has an advantage. It does not bias human performance in the target task. In this case the classification of actions has to be done using video only. Hence, as much as possible information must be extracted from the video stream in order to conduct movement analysis. The first modality is the raw information of pixel colour values. Motion is an important modality, extracted by optical flow, as investigated in [16]. It was proved to be efficient in terms of classification performance. Improvements can be achieved by making use of other information from the video recordings, such as the sound. Another possibility is to extract an information which is the interpretation result of raw data. Thus, we consider a human pose expressed via joints spatial coordinates which can be computed from the same videos. The purpose is to make cameras "smart" to analyse sport practices [7].

In this work, we investigate the use of pose information for classification inspired by the work carried out in [17, 18]. In the original twin model that takes as input RGB stream and its estimated optical flow, a third branch with pose information is added to the network. The branches are fused at the latest stage of the network through several bilinear layers. Experiments are performed on the `TTStroke-21` dataset. We solve two distinctive tasks: classification only and joint classification and segmentation from videos. Our method achieves slightly better performance on the classification task but much better scores on the joint classification and segmentation task through the use of pose and a fusion approach. We also present the opportunities that the pose offers for further movement analysis to enrich the feedback to the users.

The paper is organised as follows: computed modalities and classification method are introduced in section 2. Results and other potential applications of the pose information are discussed in section 3. Conclusion and prospects are drawn in section 4.

2 PROPOSED APPROACH

To deal with the low inter-class variability of the actions proper to fine-grained action classification, the most complete information from video must be extracted, i.e. both appearance (RGB) and motion (Optical Flow) modalities. Spatio-temporal convolutions in the network are performed on cuboids of RGB frames and on cuboids of optical flow (OF). Pose joints are also processed by temporal convolutions. All three modalities are processed simultaneously through a three-stream architecture as presented in Figure 2.

2.1 Optical Flow Estimation

As presented in [16], the OF and its normalization can strongly impact the classification results. The same motion estimator reaching best classification performances is used thereafter. The method is based on iterative re-weighted least square solver [11]. Each OF frame $V = (v_x, v_y)$ is encoded with its horizontal v_x and vertical v_y motion components being computed from two consecutive RGB frames. In order to only keep foreground motions, estimated OF is smoothed with Gaussian filter with kernel size 3×3 and multiplied using Hadamard product by the computed foreground mask M_{FG}: $V_{FG} = V \odot (M_{FG}, M_{FG})$ [38].

2.2 Region Of Interest Estimation

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

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

with parameter α set empirically to 0.6, $\Omega = (\omega_x, \omega_y) = (320, 180)$ the size of video frames. Function $f_\omega(u, S) = max(min(u, \omega - \frac{S}{2}),\ \frac{S}{2})$ allows to define ROI without the image

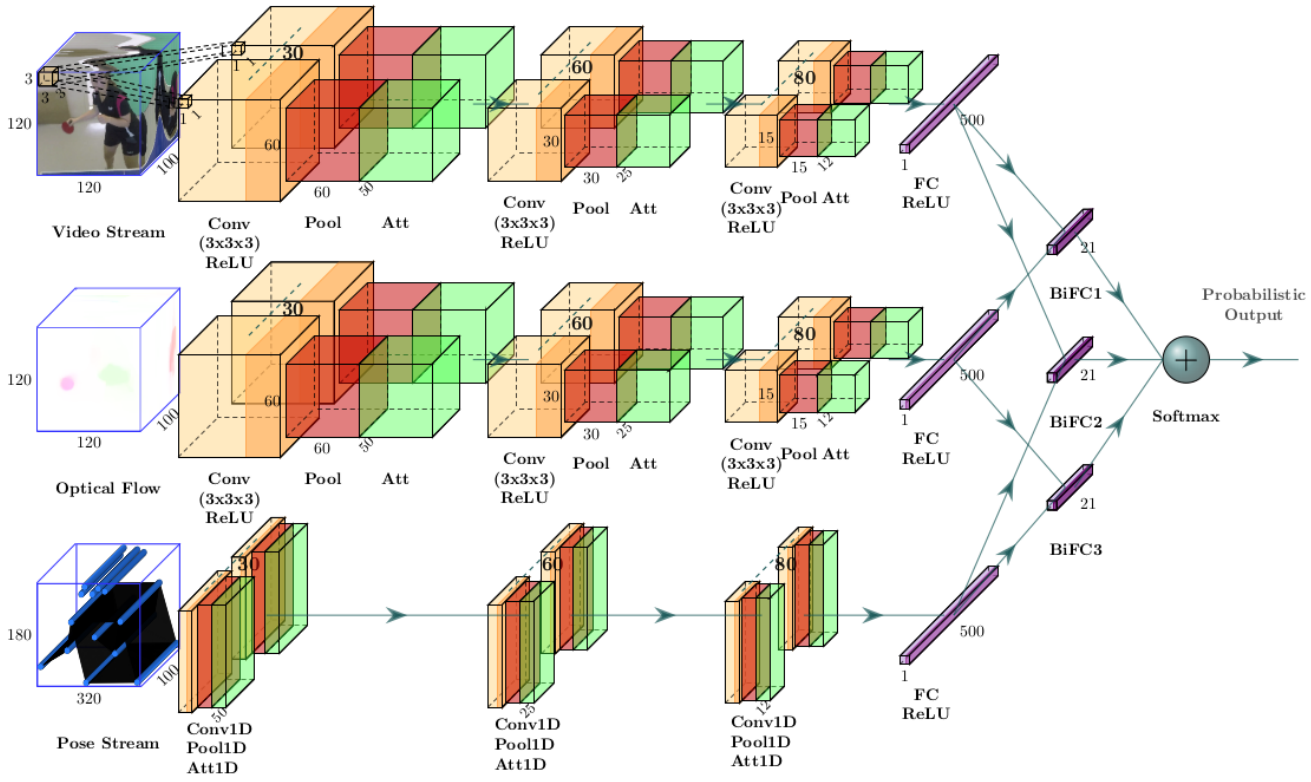


Figure 2: Three-Stream architecture processing RGB, optical flow and pose data in parallel with spatio-temporal convolutions.

border. The size of cuboids are $(W \times H \times T) = (120 \times 120 \times 100)$ which corresponds to a duration of 0.83s. To avoid jittering within the RGB and OF, a Gaussian filter of kernel size $k_{size} = 41$ ($\frac{1}{3}$ second) and scale parameter $\sigma_{blur} = 0.3 * ((k_{size} - 1) * 0.5 - 1) + 0.8$ is applied along the temporal dimension to average the ROI center position. These parameter values were chosen experimentally and are suitable for the 120 fps video frame rate.

2.3 Pose Estimation

The pose is computed from single RGB frames using the PoseNet model [21]. Its implementation is available online[1]. It supplies poses and human joints positions and their confidence score. In addition, the pose position (mean of the joint coordinates) and its attributed score are used, leading to a descriptor vector J with elements such as:

$$J(i) = (x_i, y_i, s_i)^T \tag{2}$$

with i the i^{th} joint or the pose, x_i and y_i its horizontal and vertical coordinate and s_i its associated score.

2.4 Data Normalization

To map their values into interval $[0, 1]$, RGB data are normalized by theoretical maximum, while joints position x_i and y_i are normalized with respect to the width and height of the video frames: W_{img} and H_{img}. The OF is 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 3:

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

with v and v^N representing respectively one component of the OF V_{FG} and its normalization. This normalization scales values into [-1,1] and increases the magnitude of most of vectors making the OF more relevant for classification.

2.5 Model Architecture

The architecture is inspired from the Twin Spatio-Temporal Convolutional Neural Network - T-STCNN with attention mechanisms presented in [18] which takes as input the OF and RGB values through two branches using 3D convolutions and attention mechanism. Compared to the latest, our network has three branches and takes as additional input joint coordinates. Furthermore the fusion step is adapted to fuse the three modalities.

As depicted in Figure 2, the networks perform 3D (spatio-temporal) and 1D (temporal) convolutions. The two first branches are composed of three 3D convolutional layers with 30, 60, 80 filters

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

respectively which can be described by equation 4:

$$out(j) = bias(j) + \sum_{k=0}^{C_{in}-1} weight(j,k) \star X(k) \quad (4)$$

where j is the j^{th} output channel, C_{in} the number of channels of the input X and $\star$ is the valid 3D cross-correlation operator. Each branch takes cuboids of RGB values and OF of size $(W \times H \times T)$ with respectively 3 and 2 channels. 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. The output of the successive convolutions is then flattened to feed a fully connected layer: $y = xA^T + b$ of length 500.

An extra branch processes the pose data of length T. It follows the same organization than the two other branches but uses 1D temporal convolutions over all the joints coordinates and scores (see eq. 2) at the first convolution leading to $N_{joints} \times 3$ channels. This operation is similar to equation 4 using simple cross-correlation. A max-pooling operation is performed 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 length $N_{classes}$, which represents 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 is performed on-the-fly on the train set. Each stroke sample is fed to the model once per epoch. For temporal augmentation, T successive data from the RGB, OF and Pose modalities, are extracted following a normal distribution around the center of the stroke video segment with a standard deviation of $\sigma = \frac{\Delta t - T}{N}$ with $N = 6$. Spatial augmentation is performed with random rotation in range $\pm 10^\circ$, 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 are updated accordingly. Transformations are applied on the region of interest to avoid inputting regions outside the image borders. During the test phase, no augmentation is performed and the T extracted frames are temporally centered on the stroke segment.

2.7 Training Phase

All models are trained from scratch using stochastic gradient descent with Nesterov momentum and weight decay. Cross-entropy loss is used as objective function. A learning rate scheduler is used, which reduces and increases the learning rate when the observed metric (validation loss) reaches a plateau. Warm restart technique [14] is used: weights and state of the model are saved when performing the best (lowest validation loss) and re-loaded when the learning rate is updated. This allows to re-start the optimization process from the past state with a new learning rate in the gradient descent optimizer.

The following parameters were found optimal after successive experimental trials using grid search. Grid search was used for the following parameters: $startlearningrate$, $patience$ and the number of epochs considered for comparing the training loss averages.

Training process starts with a learning rate of 0.01. A number of epochs: $patience$, set to 50, is considered before updating the learning rate, unless the performance drastically dropped (in our case: 0.7 of the best validation accuracy obtained).

The metric of interest is the training loss: if its average on the last 25 epochs is greater than its average on the 35 epochs before, the process is re-started from the past state and the learning rate divided by ten until reaching 10^{-5}. After this step, the learning rate is set back to 0.01 and the process continues. This technique differs from decreasing only by step [36] since the learning rate might re-increase if no amelioration is observed.

3 EXPERIMENTS AND RESULTS

We compare results with the original T-STCNN with attention mechanism from [18] and the Two-Stream I3D model [2], all trained and tested from scratch on `TTStroke-21` (Fig. 3), and fed with cuboids of same size. As an ablation study, the three-stream network is trained with and without attention mechanism on the third branch, using in both cases a momentum of 0.5, a weight decay of 0.05 and a batch size of 5 over 1500 epochs. The learning rate varies between 0.01 and 0.00001. The two tasks are considered: i) pure classification and ii) joint classification and segmentation. We also widen the field of application by considering the pose estimation for movement analysis.

3.1 TTStroke-21 Dataset

`TTStroke-21`, depicted in Figure 3, is composed of table tennis videos, recorded indoors at different frame rates. The players are filmed in game or training situations, performing in natural conditions without marker. The videos have been annotated by table tennis players and experts from the Faculty of Sports (STAPS) of the University of Bordeaux, France. The number of classes considered is $N_{classes} = 21$:

- **8 services:** *Serve Forehand Backspin, Serve Forehand Loop, Serve Forehand Sidespin, Serve Forehand Topspin, Serve Backhand Backspin, Serve Backhand Loop, Serve Backhand Sidespin, Serve Backhand Topspin*;
- **6 offensive strokes:** *Offensive Forehand Hit, Offensive Forehand Loop, Offensive Forehand Flip, Offensive Backhand Hit, Offensive Backhand Loop, Offensive Backhand Flip;*
- **6 defensive strokes:** *Defensive Forehand Push, Defensive Forehand Block, Defensive Forehand Backspin, Defensive Backhand Push, Defensive Backhand Block, Defensive Backhand Backspin*;

and an extra *negative* class.

In the following experiments, 129 of videos recorded at 120 fps are used. They represent a total of 1048 actions/strokes. From these temporally segmented table tennis strokes, 106 negative additional samples are extracted from the rest of the videos. A larger number of negative samples could have been extracted, but this choice was made to have a lighter class imbalance, speed up the training process and be consistent with the previous experiments. The dataset is distributed in Train, Validation and Test sets with 0.7, 0.2 and 0.1 proportions as in [17]. Extracted frames of size (1920×1080), are resized to $(W_{img} \times H_{img}) = (320 \times 180)$ before computing modalities.

(a) Acquisition Setup.

(b) Annotation Platform.

(c) Samples of a Stroke

Figure 3: TTStroke-21 Dataset.

Note that all computed human joints are not considered in pose data. Some of them might not be visible in the videos, e.g. knees and the ankles, which are often hidden by the table. We consider 13 human joints: nose, eyes, ears, shoulders, elbows, wrists and hips. This leads to a descriptor J of length $N_J = 14$. Furthermore, other players may appear in the scene, which leads to the detection of several poses in the same frame. In this case, only the closest pose from the previously computed ROI center is considered. If no pose is detected (which is the case for 25% of the frames), the descriptor vector is filled with ROI center coordinates and a score of 0. Miss-detection of the pose happens in situations when the player is out of the camera field of view. This happens when the ball leaves the table and the player has to catch it, or at the beginning and at the end of the game.

3.2 Pure Classification Task

Results are compared with different models that have been tested following an ablation method [18]. As it can be observed from table 1, the I3D model performances are worse compared to the others. The network is too deep for such a limited real-life dataset and the challenging fine-grained task. Furthermore, the classification performances of the Twin model and the Three-Stream model are similar. However, room for improvement still remains on the Three-Stream Network using attention mechanism since the gap between validation and train accuracy is lower. Convergence is achieved at epoch 1176 for the latest model after only 736 effective epochs (counting only epochs following and saving of the state). The other models reach convergence only after epoch 1400.

Preliminary results also have shown that the Pose alone could not achieve convergence and was able to classify with 22% of accuracy only on the test set for the pure classification task. The importance of the combination of different information sources is thus obvious.

3.3 Joint Classification and Segmentation Task

Similarly to [17], joint classification and segmentation of video segments is performed using an overlapping sliding window. Different decisions are investigated to flatten the obtained probabilities along the temporal dimension using a window size of 150 for "Vote" and "Avg" rules, and size 201 for "Gauss" rule. Once more, these window sizes were fixed after a preliminary grid search. The decision rules were respectively: i) majority vote, ii) average decision rule, and iii) weighted decision fusion using a Gaussian kernel. Performances are reported with all the labels, and also when the negative class is not considered. This second evaluation is motivated by the fact that most parts of a video are constituted of negative labels. Indeed, all portions between stokes are considered as negative: i.e. when the player is getting ready, when the match or training session end, and when the player is resting.

Superiority of the Three-stream network is better observed for this joint classification and detection task, see table 2. On the first part of the table, the Three-Stream Net is able to reach 63.9% of accuracy against 47.7% for the Twin model. Frame wisely, this represents a precision of 0.99 and a recall of 0.42 with regards to the negative class, leading to a F-score of 0.59. This means the model is still more likely to classify as a stroke some frames belonging to a the negative class. However this score is also biased by the frame wised approach of the evaluation. This is why the second part of the table is of better interest: the add of the third branch allows to boost the performance up to 18% compared to the model without pose information. The attention mechanism performs slightly lower, which might be overcome with a longer training phase as observed earlier. Overall, better performances are noticed for all models when not considering the negative samples, which can be

Table 1: Classification Performance in terms of Accuracy %.

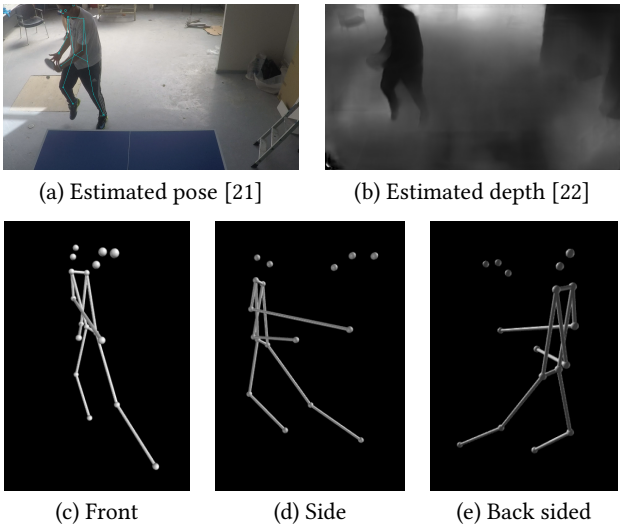
Models	Train	Validation	Test
RGB - I3D [2]	98	72.6	69.8
RGB-STCNN [16]	98.6	87	76.7
RGB-STCNN†	96.9	88.3	85.6
Flow - I3D [2]	98.9	73.5	73.3
Flow-STCNN [16]	88.5	73.5	74.1
Flow-STCNN†	96.4	83.5	79.7
RGB + Flow - I3D [2]	99.2	76.2	75.9
Twin-STCNN	99	86.1	81.9
Twin-STCNN†	97.3	87.8	87.3
Three-Stream Net.*	97	90	87.3
Three-Stream Net.†	95.8	86.5	87.3

† using attention mechanism on all branches
* using attention mechanism only on the OF and RGB branches

Table 2: Performance of Stroke Detection and Classification.

	Accuracies			
Models	Gross	Vote	Avg	Gauss
T-STCNN†	31	46.8	47.7	47.3
Three-Stream Net.*	43.6	63.1	**63.9**	62.9
Three-Stream Net.†	37.3	57.9	59	57.8
without taking into account the negative labels				
T-STCNN†	45.2	63.8	65.6	67.9
Three-Stream Net.*	69.6	83.7	84.3	**85.6**
Three-Stream Net.†	66.6	82.1	82.8	84.1

† using attention mechanism on all branches
* using attention mechanism only on the OF and RGB branches



(a) Estimated pose [21] (b) Estimated depth [22]

(c) Front (d) Side (e) Back sided

Figure 4: Combination of the pose and depth estimated from a single image (first row) to build a 3D skeleton (second row).

more challenging to classify. This may be due to all different and nonconventional gestures when a player attempts to catch a lost ball, leading to features similar to a stroke and classified as such.

3.4 Perspectives for Movement Analysis

Pose information may also be very interesting to assess the player's performance and the efficiency of his/her gesture. The organization of the joints skeleton, during a movement can be compared with a baseline to give an appreciation of the stroke performed [20]. Richer feedback could also be given by adding depth information, computed from a single image, in order to create a 3D model of the stroke. Such a representation is drawn in Figure 4. `TTStroke-21` does not offer for the moment any qualitative annotations for strokes. Such quality assessment needs to be built by experts in the field and may be one perspective of such dataset in order to widen its application.

4 CONCLUSION

We have proposed a three-stream network with different kinds of convolutions and input data: raw pixels values and optical flow undergo 3D (2D + time) convolution, while the pose-vectors are submitted to the branch with temporal convolution. Pose information, fused with RGB and optical flow branches, yields much better performances (up to 18%) in the joint classification and segmentation task. Further analysis may be conducted in this task by developing other evaluation methods not frame-wised.

Improvements can be achieved by developing a better pose estimator which can consider temporal information for avoiding misdetected poses/joints and obtain a better precision of the skeletal joints coordinates. Furthermore, pose information coupled with other technology may be one step forward to gesture analysis in order to assess athletes performance.

REFERENCES

[1] Aaron K. Baughman, Eduardo Morales, Gary Reiss, Nancy Greco, Stephen Hammer, and Shiqiang Wang. 2019. Detection of Tennis Events from Acoustic Data. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, MMSports@MM 2019, Nice, France, October 25, 2019*, Rainer Lienhart, Thomas B. Moeslund, and Hideo Saito (Eds.). ACM, 91–99. https://doi.org/10.1145/3347318.3355520

[2] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*. IEEE Computer Society, 4724–4733.

[3] Alejandro Cartas, Petia Radeva, and Mariella Dimiccoli. 2020. Activities of Daily Living Monitoring via a Wearable Camera: Toward Real-World Applications. *IEEE Access* 8 (2020), 77344–77363.

[4] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. 2018. PoTion: Pose MoTion Representation for Action Recognition. In *CVPR*. IEEE Computer Society, 7024–7033.

[5] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, François Brémond, and Gianpiero Francesca. 2019. Toyota Smarthome: Real-World Activities of Daily Living. In *ICCV*. IEEE, 833–842.

[6] Christopher J. Ebner and Rainhard Dieter Findling. 2019. Tennis Stroke Classification: Comparing Wrist and Racket as IMU Sensor Position. In *MoMM*. ACM, 74–83.

[7] 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 *WACV*. 446–455.

[8] Mehrnaz Fani, Kanav Vats, Christopher Dulhanty, David A. Clausi, and John S. Zelek. 2019. Pose-Projected Action Recognition Hourglass Network (PARHN) in Soccer. In *CRV*. IEEE, 201–208.

[9] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The AVA-Kinetics Localized Human Actions Video Dataset. *CoRR* abs/2005.00214 (2020).

[10] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. RESOUND: Towards Action Recognition Without Representation Bias. In *ECCV (6) (Lecture Notes in Computer Science, Vol. 11210)*. Springer, 520–535.

[11] Ce Liu. 2009. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[12] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. 2017. Global Context-Aware Attention LSTM Networks for 3D Action Recognition. In *CVPR*. 3671–3680.

[13] 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 *IDCS (Lecture Notes in Computer Science, Vol. 11874)*. Springer, 1–10.

[14] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR (Poster)*. OpenReview.net.

[15] 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.

[16] 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 *ICIP*. IEEE, 554–558.

[17] 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.

[18] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2021. 3D Attention Mechanism for Fine-Grained Classification of Table Tennis Strokes using a Twin Spatio-Temporal Convolutional Neural Networks. In *ICPR*. IEEE Computer Society.

[19] Alaeddine Mihoub, Gérard Bailly, Christian Wolf, and Frédéric Elisei. 2016. Graphical models for social behavior modeling in face-to face interaction. *Pattern Recognit. Lett.* 74 (2016), 82–89.

[20] Marion Morel, Catherine Achard, Richard Kulpa, and Séverine Dubuisson. 2017. Automatic evaluation of sports motion: A generic computation of spatial and temporal errors. *Image Vis. Comput.* 64 (2017), 67–78.

[21] 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 *ECCV (14) (Lecture Notes in Computer Science, Vol. 11218)*. Springer, 282–299.

[22] Michaël Ramamonjisoa and Vincent Lepetit. 2019. SharpNet: Fast and Accurate Recovery of Occluding Contours in Monocular Depth Estimation. In *ICCV Workshops*. IEEE, 2109–2118.

[23] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2017. LCR-Net: Localization-Classification-Regression for Human Pose. In *CVPR*. IEEE Computer Society, 1216–1224.

[24] 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.

[25] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding. In *CVPR*. IEEE, 2613–2622.

[26] Tomohiro Shimizu, Ryo Hachiuma, Hideo Saito, Takashi Yoshikawa, and Chonho Lee. 2019. Prediction of Future Shot Direction using Pose and Position of Tennis Player. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, MMSports@MM 2019, Nice, France, October 25, 2019*, Rainer Lienhart, Thomas B. Moeslund, and Hideo Saito (Eds.). ACM, 59–66. https://doi.org/10.1145/3347318.3355523

[27] Bharat Singh, Tim K. Marks, Michael J. Jones, Oncel Tuzel, and Ming Shao. 2016. A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. In *CVPR*. IEEE Computer Society, 1961–1970.

[28] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. 2020. A Short Note on the Kinetics-700-2020 Human Action Dataset. *CoRR* abs/2010.10864 (2020).

[29] Khurram Soomro, Haroon Idrees, and Mubarak Shah. 2019. Online Localization and Prediction of Actions and Interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 459–472.

[30] 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).

[31] Roman Voeikov, Nikolay Falaleev, and Ruslan Baikulov. 2020. TTNet: Real-time temporal and spatial video analysis of table tennis. (2020), 3866–3874.

[32] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. 2016. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 8 (2016), 1583–1597.

[33] Erwin Wu and Hideki Koike. 2020. FuturePong: Real-time Table Tennis Trajectory Forecasting using Pose Prediction Network. In *CHI Extended Abstracts*. ACM, 1–8.

[34] 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.

[35] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. 2019. PA3D: Pose-Action 3D Machine for Video Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 7922–7931. https://doi.org/10.1109/CVPR.2019.00811

[36] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. In *BMVC*. BMVA Press.

[37] Tianhang Zheng, Sheng Liu, Changyou Chen, Junsong Yuan, Baochun Li, and Kui Ren. 2020. Towards Understanding the Adversarial Vulnerability of Skeleton-based Action Recognition. *CoRR* abs/2005.07151 (2020).

[38] 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.