Optimal Choice of Motion Estimation Methods for Fine-Grained Action Classification with 3D Convolutional Networks

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

OPTIMAL CHOICE OF MOTION ESTIMATION METHODS FOR FINE-GRAINED ACTION CLASSIFICATION WITH 3D CONVOLUTIONAL NETWORKS

Pierre-Etienne Martin[1], *Jenny Benois-Pineau*[1], *Renaud Péteri*[2] *and Julien Morlier*[3]

[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France
[2]MIA, La Rochelle University, La Rochelle, France
[3]IMS, University of Bordeaux, Talence, France

ABSTRACT

Detecting and classifying human actions in videos is one of the current challenges in visual content analysis and mining. This paper presents a method for performing a fine-grained classification of sport actions using a Siamese Spatio-Temporal Convolutional Neural Network (SSTCNN) model. This model takes RGB images and Optical Flow field as input data. Our first contribution is the comparison of different Optical flow methods and a study of their influence on the classification score. We also present different normalization methods for the optical flow that drastically impact results, boosting performances from 44% to 74% of accuracy. Our second contribution is the detection and classification of actions in videos performed using a sliding temporal window. It leads to a satisfying score of 81.3% over the whole dataset `TTStroke-21`.

***Index Terms*—** Deep Learning, Optical Flow, Data Normalization, Action classification, Spatio-temporal convolution

1. INTRODUCTION

Action detection and classification is one of the main challenges in visual content analysis and mining. Sport video analysis has been in the past years a very popular research topic, due to the variety of application areas, ranging from multimedia intelligent devices with user-tailored digests [1], up to analysis of athletes' performances [2]. Datasets focused on sports activities [3] or including a large amount of sport activity classes [4, 5] are now available with many researches benchmarking on those datasets [6, 7, 8]. A large amount of works is also devoted to fine-grained classification through the analysis of sport gestures using motion capture systems [9]. However, body-worn sensors and markers can disturb the natural behavior of sportsmen. Furthermore, motion capture devices are not always available for potential users, be it a University Faculty or a local sport team. Our work is focused on detection and recognition of strokes in table tennis. It is a first step in a wide research program which goal is to give tools for sport coaches to improve performances of young athletes using recorded videos of training and playing sessions. In order to reach the largest audience, recordings have to be performed by widespread and cheap video cameras, e.g. GoPro. We use a dataset specifically recorded in a sport faculty facility and continuously completed by students and teachers. This dataset is constituted of player-centred videos recorded in natural conditions without markers or sensors. It comprises 20 table tennis strokes and a rejection class. The problem is hence a typical research topic in the field of video indexing: for a given recording, we need to label the video by recognizing and temporally segmenting each stroke appearing in the whole video. Motion information is obviously a crucial clue for recognizing table tennis strokes. Review of the state-of-the art shows that spatial features from RGB images are also needed to attain reasonable accuracies for action classification, be it in sport [10] or in specific cultural content [11]. Hence, it is necessary to use both multi-modal data and temporal information. Contrary to [12, 13] which use one channel of a 3D tensors for encoding temporal information, we are using 3D convolutions of video frames, similarly to the promising results obtained in [14, 15, 7]. However, to efficiently fuse data of limited spatial localization of moving sportsmen in image plane, variable magnitude of motion in each stroke, an adequate normalization of motion data has to be elaborated. Our first contribution is thus a comparison of different optical flow (OF) estimation methods and their normalization and how they influence the training of a long term convolutional neural network such as the one proposed by [10]. Our second contribution is the detection, classification and temporal segmentation of table tennis strokes in videos, based on temporal sliding windows and a spatio-temporal CNN introduced in [16].

The remainder of the paper is organized as follows: section 2 presents our model and the classification method. In section 3, different OF methods are compared on Sintel benchmark [17] with different normalization approaches. Results and performance assessment for classification and recognition are presented in section 4. Conclusion and prospects are drawn in section 5.

This work was supported by Region Nouvelle Aquitaine (grant CRISP) and Bordeaux Idex Initiative

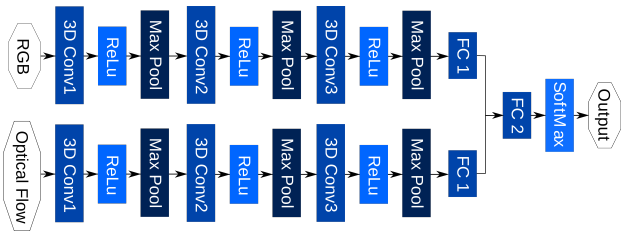

Fig. 1. Our Siamese (SSTC) architecture.

2. PROPOSED METHOD

Classification of actions is performed in video scenes of a single table tennis player performing a series of strokes. To limit analysis area in full HD video frames, we resize them to 320×180 pixels and compute their OF offline. A spatial region-of-interest (ROI) of size ($W \times H$) is then extracted, as described in [16], based on the foreground [18] OF values but using dilation of a mask and smoothed OF values. Our model is feed with 3D tensors of size ($W \times H \times T$) with RGB images and/or OF data. The model is tested for classification and detection through classification on dataset `TTStroke-21`.

2.1. Architecture

The Siamese Spatio-Temporal Convolution network (SSTC), is constituted of 2 individual branches with three 3D convolutional layers with 30, 60, 80 filter response maps, followed by a fully connected layer of size 500 (Fig. 1). One branch takes RGB values as input, and input of the other branch is the OF preliminary estimated on the current video frame $\mathbf{V} = (v_x, v_y)$. The 3D convolutionnal layers use $3 \times 3 \times 3$ space-time filters with a dense stride and padding set to 1 in each direction. The two branches are fused through a final fully connected layer of size 21 followed by a Softmax function for outputting a classification score. The architecture of the branches, constituted of 3 max pooling layers, has a similar depth as the AlexNet architecture [19]. The latter has proven its efficiency for image classification still remaining sufficiently shallow for being adapted to low-resolution videos.

2.2. Model Training

We train three network models: '`RGB`', '`Optical Flow`' and '`Siamese`'. The '`Siamese`' model uses the full architecture presented in section 2.1 while '`RGB`' and '`Optical Flow`' models are constituted of one branch only. The optimization method is the popular stochastic gradient descent with Nesterov momentum used in [19]. We have chosen a constant learning rate, set to 0.01 for '`RGB`' and '`Optical Flow`' models and to 0.001 for the '`Siamese`' model. The other parameters are momentum of 0.5, weight decays of 0.005, and a batch size of 10. Most of the trainings were done with 500 epochs but for the sake of comparison we trained some models with 1500 epochs.

2.3. Detection through Classification

To detect actions in videos, we classify overlapping temporal sliding windows of size T. A vector of probability scores P of size $T_{video} - T$ is obtained, with T_{video} being the length of the video. To avoid border effects when classifying the whole video, we extrapolate P by simple copy of the first and last probability score respectively at the beginning and at the end of our probability vector. Different rules were used for classifying each frame using a decision window WD. Our two first decision rules are the majority vote and the average probability. Our third decision rule is based on filtered probability using a temporal Gaussian kernel.

3. OPTIMAL OPTICAL FLOW SELECTION AND NORMALIZATION

Optical flow estimation is of primarily importance in the analysis of spatio-temporal actions. In this work we consider the following OF methods : Farneback [20], Beyond Pixel (BP) [21] used in [16], DIS [22], TVL1 [23] and DeepFlow [24]. The quality of each method is evaluated with usual metrics such as Mean Squared Error (MSE) of motion compensation. We compute average MSE (aMSE) on video sequences, namely on a strong-motion sequence of an Offensive Forehand Hit stroke from `TTStroke-21` dataset (240 frames) denoted as FH in Table 1. The popular Sintel benchmark [17] is also used. This is a dataset of synthetic videos with available reference optical flows, with some sequences with strong aliasing effects and random texture. In this case the average angular and end-point errors are computed with regard to the ground truth denoted as aAE and aEPE respectively. It is indeed well-known that OF methods may have difficulties on flat areas and can give noisy results on highly contrasted borders due to aliasing effects. Thus, a better motion compensation does not yield better optical flow estimation. As several kinds of assessment are necessary, different normalization methods are considered according to the computed OF.

3.1. Selection of Optical Flow estimator

The BP method [21] used in our previous work [16] does not perform well on Sintel Benchmark as it is very sensitive to random textures. According to aMSE values obtained on FH stroke sequence (see Table 1), the best method is the Dense Inverse Search (DIS) with spatial propagation with preset parameters denoted '`Medium`' in OpenCV [22]. DIS method is focused on reducing time complexity but still yields competitive accuracy. However, aMSE is not a good metric for evaluating an OF estimator due to aliasing and texture effects, occlusions and illumination variations. Hence aMSE value on Sintel Benchmark computed with available ground truth OF is higher than that one supplied by the most of OF estimators. Its standard deviation is also very high showing the complexity

Table 1. Optical Flow method comparison

	Sintel Benchmark			FH
	aEPE	aAE	aMSE	aMSE
Frame Diff	-	-	872 ± 1017	33 ± 12
Ground Truth	-	-	407 ± 778	-
Farneback	11 ± 18	$.7 \pm .33$	365 ± 771	21 ± 7.8
BP	6.4 ± 13	$.42 \pm .27$	316 ± 628	20 ± 3.9
DeepFlow	6.8 ± 15	$.37 \pm .26$	384 ± 808	24 ± 5.8
DeepFlow with matches	**2.6±5.7**	**.3±.18**	348 ± 711	25 ± 5.2
TVL1	9.3 ± 17	$.54 \pm .32$	423 ± 752	20 ± 4.1
DIS Medium	5.5 ± 11	$.46 \pm .29$	**290±540**	20 ± 3.9
DIS Medium†	4.8 ± 9.8	$.44 \pm .26$	318 ± 670	**20±4.6**
DIS Medium*	4.8 ± 9.8	$.43 \pm .27$	297 ± 559	20 ± 3.9
DIS Medium†*	4.8 ± 9.8	$.43 \pm 2.7$	297 ± 559	20 ± 3.9
DIS Fast	6 ± 12	$.52 \pm .31$	299 ± 526	24 ± 5.7
DIS Fast†	5.3 ± 11	$.49 \pm .28$	312 ± 605	24 ± 5.6
DIS Fast*	5.2 ± 9.8	$.48 \pm .28$	295 ± 523	24 ± 5.6
DIS Fast†*	4.9 ± 12	$.48 \pm .27$	321 ± 655	24 ± 5.6
DIS Ultrafast	7.1 ± 13	$.58 \pm .32$	307 ± 536	23 ± 5.6
DIS Ultrafast†	5.6 ± 11	$.53 \pm .28$	314 ± 601	23 ± 5.3
DIS Ultrafast*	5.9 ± 11	$.53 \pm .29$	297 ± 513	23 ± 5.4
DIS Ultrafast†*	5.2 ± 11	$.51 \pm .27$	323 ± 654	23 ± 5.3

* : with Temporal propagation † : with Spatial propagation

of the Sintel dataset and limitation of aMSE metric. According to average angular and average end-point errors, the best estimator is DeepFlow, a variational approach for optical flow combined with matching algorithm [24]. As illustrated in Fig. 2, DeepFlow method does not generate false movements on flat regions contrary to DIS Medium estimator with spatial propagation which performs the best with respect to MSE. However computation time is hundred times faster with DIS estimator than with DeepFLow or BP estimators. However, computation time is not here a crucial issue, because we are not interested in online predictions but in obtaining an accurate OF estimation. Hence in our work we will use DeepFlow and BP as it has shown good results in terms of classification accuracy on our dataset in [16].

3.2. Normalization

RGB image channels are normalized by their theoretical maximum value (255 in our case) to map them into interval [0,1]. For the Optical Flow $\mathbf{V} = (v_x, v_y)$, different approaches are possible. Three methods have been tried: the first one is to normalize each component of $\mathbf{V}$ by its maximum absolute value over the whole dataset. We reference this method as 'MAX'. The second method is to normalize each component of $\mathbf{V}$ by the mean μ and the standard deviation σ of the distribution over the whole dataset of frame maximum absolute values. We reference this method as 'NORMAL' (Eq. 1). In the two following equations v and v^N represent respectively one component of the OF $\mathbf{V}$ and its normalization.

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

a. RGB b. DIS c. DeepFlow

Fig. 2. Optical Flow comparison

The third method, denoted 'LOG', is similar to the previous one but takes the logarithm of the distribution over the whole dataset of frame maximum absolute values of a component (Eq. 2).

$$v_{log} = log(|v| + 1)$$
$$v' = \frac{v_{log} - (\mu_{log} - 3\times\sigma_{log})}{6\times\sigma_{log}}$$
$$v_n(i,j) = \begin{cases} 0 & \text{if } v'(i,j) \leq 0 \\ SIGN(v(i,j)) \times v'(i,j) & \text{if } 0 < v'(i,j) < 1 \\ SIGN(v(i,j)) & \text{otherwise.} \end{cases} \quad (2)$$

Obviously, the MAX method strongly reduces the magnitude of most motion vectors. However the 'NORMAL' normalization method increases the magnitude of most vectors, while 'LOG' flattens the value distribution.

3.3. Data Augmentation

Data augmentation is a necessary step for training Deep NNs. It is performed on the fly to save storage space. For spatial augmentation we apply random rotation with angle θ in the range $\pm 10°$, a random translation (t_x, t_y) in the range ± 0.1 in x and y directions, and a random homothety k in range 1 ± 0.1 both on RGB images and optical flow. For the latter, rotation is performed such as $R(\theta)V$ with $R(\theta)$ being the rotation matrix of angle θ. Transformations are applied with respect to the center of the ROI. Finally we perform horizontal flip with probability of 0.5. Note that the flip on optical flow means also changing the sign of v_x.

4. EXPERIMENTS AND RESULTS

The mean duration of strokes in TTStroke-21 is 174 ± 43.14 frames with a minimum of 99 and a maximum of 276 frames. To avoid having two full strokes in a same temporal window, we set T to 100 frames, which represents 0.83 seconds. Our model is then fed with cuboids of videos of size $(W \times H \times T) = (120 \times 120 \times 100)$ extracted from RGB images and/or the OF. We performed the evaluation of the normalization method and the classification in video on a specifically recorded dataset TTStroke-21 which has been introduced in [16]. To our best knowledge, this table tennis dataset of training athletes is a unique one. We do not tackle the problem of action recognition in sport video in general, but focus on fine grained stroke recognition. A protocol can be inspired from this work for other sports, but cannot be applied directly to another sport dataset. Indeed, each sport imposes its own constraints and it may be very perilous to generalize a method

Table 2. Performances of the Optical Flow normalizations

	Accuracies				
Normalization	**Train**	**Val**	**Test**	**T.Vote**	**T.Avg**
BP					
MAX	53.5	44.4	43.1	44	44
NORMAL	88.5	73.5	69.8	72.4	**74.1**
LOG	97.8	75.7	65.5	66.4	68.1
DeepFlow					
MAX	38.5	36.5	25	28.5	27.6
NORMAL	34	35.7	25.9	25.9	26.7
LOG	45.3	37	35.3	40.5	41.4

Table 3. Performances on pure classification task

	Accuracies				
Models	**Train**	**Val**	**Test**	**T.Vote**	**T.Avg**
RGB	97.7	75.7	70.7	68.1	69.8
Optical Flow	91	78.7	67.2	71.6	70.7
Siamese	86.8	62.6	54.3	52.6	56.9
with data augmentation					
RGB	98.6	87	70.7	75.9	76.7
Optical Flow	88.5	73.5	69.8	72.4	*74.1*
Siamese	89.4	79.1	69	71.6	72.4

Table 4. Performance classification and detection task

	Accuracies		
Models	**Vote**	**Average**	**Gaussian**
RGB	72.9	74.7	74.4
Optical Flow	76.6	78.4	77.9
Siamese	80.2	81.3	**81.3**

from one sport to another. The camera shooting, the dynamic of actions, the presence of sport equipment (racket, table tennis ball, soccer ball and so on) make methods context dependent and thus not generic.

4.1. Influence of normalization method

Table 2 shows the performances of CNN classifier trained with 500 epochs using BP and DeepFlow OF estimators with the three normalization methods. It is a pure classification task, as temporal borders of each action are known. As we can see, best performances are obtained using the 'NORMAL' method with BP estimator. As explained in section 3.2, the 'NORMAL' method increases low values but does not distort them as 'LOG' method does. It is also a way to get rid of the noise from the OF computation which leads to very high values and makes the 'MAX' method less efficient. However our model trained with OF from DeepFlow estimator did not perform well. It underlines the importance of the normalization method. According to these results we decided to use the 'NORMAL' normalization with BP estimator for the next trainings.

4.2. Data Augmentation

Tables 3 and 4 show that data augmentation is helpful for both classification only and classification plus detection tasks. In table 3 accuracies are higher with data augmentation, but RGB data only remain more appropriate for the pure classification task. Table 4 shows results for classification plus detection task. Best results are obtained with our Siamese network for a sufficient number of epochs (1500). These results are obtained with BF OF method [16], 'NORMAL' normalization and designed data augmentation.

4.3. Detection and classification

To evaluate the performances of our method for detection and classification actions in videos, we compare predictions with the ground truth built from the crowdsourced annotations of the `TTStroke-21` dataset. Experiments have been conducted with the whole dataset which incorporates strokes and negative samples in the training, validation and test sets. Each model classifies the entire video with a temporal sliding window $T = 100$ frames with step 1. Majority vote and average probability method use a sliding decision window $WD = 1.5T = 150$ also with step 1. For the decision rule based on a Gaussian filter, a kernel of size $2T + 1 = 201$ and a scale parameter $\sigma = 0.5T = 50$ were used. As the detection may not be always precise in time because of errors in the crowdsourced annotations, a prediction is considered correct at the boundaries of actions (between two classes) if one of these classes is found. The maximum possible class overlap is set to 20% of the current stroke duration. Results are shown in table 4.

5. CONCLUSION AND PERSPECTIVES

In this paper we have compared several optical flow methods in terms of aMSE, aAE and aEPE metrics for selecting the optimal one as input data in a 3D CNN for fine-grained sport action classification. We have proposed three normalization schemes and studied their influence with respect to the classification accuracy. Our choices improved results up to 74.1% with models trained in a reasonable number of epochs. We also showed that for both pure classification and classification with detection, proposed data augmentation was useful for all kinds of data : RGB, OF on single branch or Siamese NN. The siamese model reached a satisfying score of 81.3% on a challenging natural dataset of Table Tennis strokes. In addition, we believe that better results can be obtained since a continuous improvement has been noticed for a greater number of epochs when training the Siamese model. The dataset `TTStroke-21` used for these experiments is still being enriched for further research and applications.

6. REFERENCES

[1] Ewa Kijak, Guillaume Gravier, Lionel Oisel, and Patrick Gros, "Audiovisual integration for tennis broadcast structuring," *Multimedia Tools Appl.*, vol. 30, no. 3, pp. 289–311, 2006.

[2] Moritz Einfalt, Dan Zecha, and Rainer Lienhart, "Activity-conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming," in *WACV*, 2018, pp. 446–455.

[3] Juan Carlos Niebles, Chih-Wei Chen, and Fei-Fei Li, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV 2010*, 2010, pp. 392–405.

[4] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. 1212.0402, 2012.

[5] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[6] Cyrille Beaudry, Renaud Péteri, and Laurent Mascarilla, "An efficient and sparse approach for large scale human action recognition in videos," *Mach. Vis. Appl.*, vol. 27, no. 4, pp. 529–543, 2016.

[7] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," *CoRR*, vol. abs/1705.07750, 2017.

[8] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid, "Potion: Pose motion representation for action recognition," in *CVPR 2018, 2018*. 2018, pp. 7024–7033, IEEE Computer Society.

[9] S. Noiumkar and S. Tirakoat, "Use of optical motion capture in sports science: A case study of golf swing," in *ICICM*, 2013, pp. 310–313.

[10] Gül Varol, Ivan Laptev, and Cordelia Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, 2018.

[11] Andrei Stoian, Marin Ferecatu, Jenny Benois-Pineau, and Michel Crucianu, "Fast action localization in large-scale video archives," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 26, no. 10, pp. 1917–1930, 2016.

[12] Marjaneh Safaei, Pooyan Balouchian, and Hassan Foroosh, "TICNN: A hierarchical deep learning framework for still image action recognition using temporal image prediction," in *IEEE ICIP*, 2018, pp. 3463–3467.

[13] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi, "Action recognition with dynamic image networks," *CoRR*, vol. abs/1612.00738, 2016.

[14] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE ICCV*, 2015, pp. 4489–4497.

[15] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE CVPR 2016*, 2016, pp. 1933–1941.

[16] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier, "Sport action recognition with siamese spatio-temporal cnns: Application to table tennis," in *CBMI 2018*. 2018, pp. 1–6, IEEE.

[17] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *IEEE ECCV*, 2012.

[18] Zoran Zivkovic and Ferdinand van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS, Lake Tahoe, Nevada, United States.*, 2012, pp. 1106–1114.

[20] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *SCIA*, Josef Bigün and Tomas Gustavsson, Eds. 2003, vol. 2749 of *LNCS*, pp. 363–370, Springer.

[21] Ce Liu, *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*, Ph.D. thesis, Massachusetts Institute of Technology, 5 2009.

[22] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool, "Fast optical flow using dense inverse search," in *ECCV*. 2016, vol. 9908 of *LNCS*, pp. 471–488, Springer.

[23] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime TV-L^1 optical flow," in *29th DAGM*, Fred A. Hamprecht, Christoph Schnörr, and Bernd Jähne, Eds. 2007, vol. 4713 of *LNCS*, pp. 214–223, Springer.

[24] Philippe Weinzaepfel, Jérôme Revaud, Zaïd Harchaoui, and Cordelia Schmid, "Deepflow: Large displacement optical flow with deep matching," in *IEEE ICCV*, 2013, pp. 1385–1392.