Optical Flow Singularities for Sports Video Annotation: Detection 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

Over the past few years, Action Recognition task has drawn considerable interests, leading to intensive researches. This is mainly due to the variety of related applications, from autonomous car to human behavior analysis.

Up to now, most of researches aim to identify various sport actions such as UCF-101 dataset[11], but, due to the exponential number of online videos and the necessity to be more and more accurate, the need of finer analysis arises.

In this working note, results for the MediaEval 2019 Sports Video Annotation "Detection of Strokes in Table Tennis" task [9] are presented. As in sport videos displacement flow appears to be one of the most useful information for stroke identification, especially to differentiate quite similar strokes, this proposal relies on a combination of spatial information and Optical Flow's singularities identification. As a result, most relevant regions of video frames for the classification task are detected.

Figure 1: Extracted Optical Flow using PWC-Net

1 INTRODUCTION

The selected task requires to analyze a single sport, which means that the analysis has to be even more precise than high inter-class variance datasets. The dataset, aiming at representing real-life sportsman training situations, is made up of videos recorded using standard cameras with unbalanced number of training samples for each stroke. No depth maps or data issued from motion capture suits are available.

This working note provides a description of the methods proposed by the team MIA on this task. Only handcrafted features extracted from video frames and optical flow are used: Histogram of oriented Gradients (HoG)[6] features and dense Optical Flow singularities's coefficients projected on Legendre basis. These features are represented by a Bag-of-Words model and the final classification is obtained by mean of a linear SVM.

2 OUR APPROACH

The great success and popularity of Deep Learning methods for 2D images recognition tasks, led many researchers to adapt these architectures to video analysis using 3D filters instead of 2D filters commonly known as 3DCNN[13].

For both manual and deep learning methods, the Optical Flow was also proved relevant, with the arrival of two-stream network architectures[10] or Siamese Network[8]. Because the automatically calculated filters of deep-learning methods could have no real human meaning compared to handcrafted approaches, we decided to extract interesting regions around the player based only on the optical flow's singularities [1–3] and did complementary analysis on this areas.

As already said, the proposed approach relies on dense accurate Optical Flow. Nowadays, one of the most popular method is probably the Farneback [7] method which starts by generating an image pyramid of different resolutions, and uses polynomial expansion to match the pixel from one resolution to another. The main issue with this method is that when an object of uniform color is moving, only the borders of that object are detected. Using Farneback provides good edges, but empty objects.

More recent methods are trying to overcome this drawback, especially, the PWC-Network [12] that use CNN pyramidal feature extraction, warping layers, and cost volume layers to match features of the first image and warped features of the second one. Our method uses such a network pre-trained using the Sintel dataset [4], an open source animated short film, to give clean boundaries like in Figure 1. Compared to the Sintel dataset, the task dataset presents a lot of compression artifacts, consequently, Gaussian blur is applied before Optical Flow extraction, and frames are resized to speed up consequent processing.

2.1 Optical Flow Singularities

Given the horizontal and vertical components U and V of the optical flow, regions of high rotation or divergence are detected by the following stage. For each frame, using a sliding window, the optical flow is locally approximated using a Legendre polynomial basis. The polynomial basis P is defined as:
$P_{K,L}(x_1, x_2) = \sum_{k=0}^{K} \sum_{l=0}^{L} x_1^k x_2^l$

To obtain precise results, a small sliding window of 50 pixels is chosen. The resulting computational cost is therefore limited as a one-dimensional polynomial basis is precise enough in such a case

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

Table 1: Global accuracy

Method	Train set	Test set
Unbalanced SVM	153/754	25/354
Position + Unbalanced SVM	426/754	46/354
Position + HoG + Unbalanced SVM	524/754	46/354
Position + Hog + Balanced SVM	485/754	50/354

:

$$U = u_{0,0}P_{0,0} + u_{0,1}P_{0,1} + u_{1,0}P_{1,0}$$
$$V = v_{0,0}P_{0,0} + v_{0,1}P_{0,1} + v_{1,0}P_{1,0}$$

After the projection, the two components are efficiently calculated on a canonical basis by approximating U and V flows as follows :

$$\begin{pmatrix} U \\ V \end{pmatrix} \simeq A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + b = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + b_1 \\ a_{21}x_1 + a_{22}x_2 + b_2 \end{pmatrix}$$

Each pixel region is then represented by a $2x2$ matrix made of canonical projection coefficients of the flow. Significant region are selected by a simple threshold:
$\Delta(A) = tr(A)^2 - 4 * det(A), \Delta(A) < 0.05$

2.2 BoW and SVM for Action Recognition

The classification task follow the Bag of Word (BoW) approach: K-Means are used to classify the various singularities (each singularity being originally represented by the four projection coefficients) into six clusters.

Except for the first run, the relative spatial positions of the singularities in the frames are also used. The frames are divided in four-squared grids and the number of singularities on each of these four regions are analysed.

For the last two runs, HoG Features, as represented by a height bins BoW, are also used but only on regions where significant singularities have been selected. This aims at quantifying the relative importance of optical flow-based and gradient based features.

As a result, each stroke is represented by an histogram with at most 18 bins (6 singularities, 8 HoG, and 4 spatial regions).

Classification is done by a cross-validated linear SVM[5], thus avoiding overfitting.

The given dataset being seriously unbalanced, a balanced SVM is used on the last run, giving penalties for the most common classes, to increate the retention rate of rare strokes.

3 RESULTS AND ANALYSIS

The proposed method leads to four runs, using only singularities for the first one, and adding additional information like HoG or the position of the singularities region for the others. The accuracy of the four runs are presented in Table 1 for both training and testing set.

The last three runs with the singularities and spatial/pixel information have pretty similar results for the test set, but the run using only the projection coefficients gives a lower global accuracy. That proves that using movement-based analyze, without using other data is not sufficient to have a good enough interpretation of a stroke, and focusing only on the flow information results in high information loss.

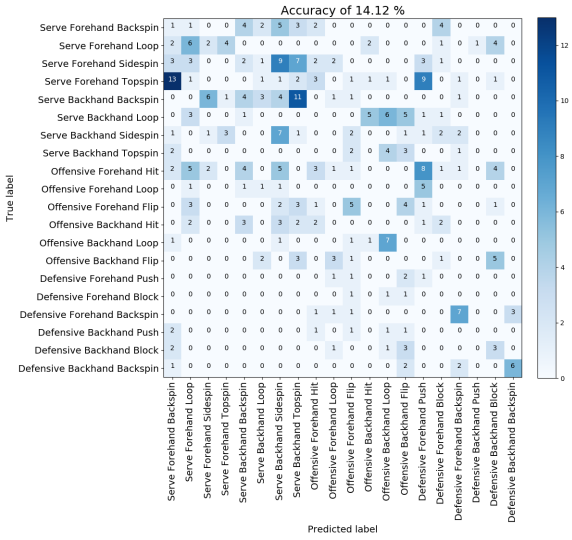

Figure 2: Accuracy of the predicted classes

The second and third run, with singularity positions and unbalanced SVM have similar results both in terms of overall accuracy and predicted classes. This behavior is unexpected as one of the run uses Hog features, while the others does not. Maybe, because only one sport is present in the dataset, the players edges are not sufficient to differentiate strokes. We used HoG on each frame, knowing that one frame alone isn't enough to know what stroke class it belongs to. We stacked them over the whole sequence without taking into account the temporal data, and that's probably why the HoG have no impact on the results overall.

On the other hand, the only run with balanced SVM provides a better overall accuracy. As said in the introduction, the dataset is heterogeneously balanced. Standard unbalanced SVM predicts the classes to increase the overall result. On this dataset, it overpredicts the most frequent classes. By using weights, balanced SVM increases its accuracy on the rare classes, resulting in a worst overall result, but in better results on rare classes.

4 DISCUSSION AND OUTLOOK

This paper presents an approach for the Sports Video Annotation on single-sport dataset task. Due to the difficulty of the task, the rare classes samples, missing metadata about right or left handed players, and different camera viewpoints, didn't achieved high performance scores, but it gives an insight of what is missing in the proposed Optical Flow's Singularities features.

There is a still rooms for improvement, mostly due to the lack of long term temporal information and the variations between two optical flows of the same stroke class when recorded by cameras on different viewpoints.

REFERENCES

[1] Cyrille Beaudry, Renaud Péteri, and Laurent Mascarilla. 2014. Action recognition in videos using frequency analysis of critical point trajectories. *2014 IEEE International Conference on Image Processing, ICIP 2014*. https://doi.org/10.1109/ICIP.2014.7025289

[2] Cyrille Beaudry, Renaud Péteri, and Laurent Mascarilla. 2016. An efficient and sparse approach for large scale human action recognition in videos. *Machine Vision and Applications* 27, 4 (2016), 529–543.

[3] Katy Blanc, Diane Lingrand, and Frédéric Precioso. 2017. SINGLETS: Multi-Resolution Motion Singularities for Soccer Video Abstraction. In *Workshop CVsports (in conjunction with CVPR) (Proceedings of the Workshop CVsports (in conjunction with CVPR))*. Honolulu (Hawaii), United States. https://hal.archives-ouvertes.fr/hal-01540342

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. 2012. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV) (Part IV, LNCS 7577)*, A. Fitzgibbon et al. (Eds.) (Ed.). Springer-Verlag, 611–625.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 886–893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[7] Gunnar Farnebäck. 2003. Two-frame Motion Estimation Based on Polynomial Expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA'03)*. Springer-Verlag, Berlin, Heidelberg, 363–370. http://dl.acm.org/citation.cfm?id=1763974.1764031

[8] P. Martin, J. Benois-Pineau, R. Péteri, and J. Morlier. 2018. Sport Action Recognition with Siamese Spatio-Temporal CNNs: Application to Table Tennis. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI 2018)*. 1–6. https://doi.org/10.1109/CBMI.2018.8516488

[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. *Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-29 October 2019*.

[10] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. *CoRR* abs/1406.2199 (2014). arXiv:1406.2199 http://arxiv.org/abs/1406.2199

[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. 2017. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. *CoRR* abs/1709.02371 (2017). arXiv:1709.02371 http://arxiv.org/abs/1709.02371

[13] Shuiwang Ji ; Wei Xu ; Ming Yang ; Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (Jan 2013), 221–231. https://doi.org/10.1109/TPAMI.2012.59