A REVISIT OF ACTION DETECTION USING IMPROVED TRAJECTORIES

Konstantinos Papadopoulos, Michel Antunes, Djamila Aouada, Björn Ottersten

Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg, Luxembourg
{konstantinos.papadopoulos, djamila.aouada, bjorn.ottersten}@uni.lu,
michel.gon.antunes@gmail.com

ABSTRACT

In this paper, we revisit trajectory-based action detection in a potent and non-uniform way. Improved trajectories have been proven to be an effective model for motion description in action recognition. In temporal action localization, however, this approach is not efficiently exploited. Trajectory features extracted from uniform video segments result in significant performance degradation due to two reasons: (a) during uniform segmentation, a significant amount of noise is often added to the main action and (b) partial actions can have negative impact in classifier's performance. Since uniform video segmentation seems to be insufficient for this task, we propose a two-step supervised non-uniform segmentation, performed in an online manner. Action proposals are generated using either 2D or 3D data, therefore action classification can be directly performed on them using the standard improved trajectories approach. We experimentally compare our method with other approaches and we show improved performance on a challenging online action detection dataset.

Index Terms— Action detection, improved trajectories, action proposals

1. INTRODUCTION

Human action detection has drawn significant attention over the past years. This active research topic of computer vision finds applications in various fields, such as video surveillance, healthcare and human-computer interaction. Still, large background data variations, inaccurate detection of starting and ending points of action and observation of partial actions [1] are challenges that need to be addressed.

There has been a substantial amount of work in the field of temporal action detection. Huang et. al. in [2] suggested a model which evaluates and discards action classes by observing partial events. In addition, in [3] a similar early event detector of short video segments was developed, in which the labels of the expected actions are provided. In [4], authors proposed the segmentation of videos into a sequence of atomic action units. Moreover, a model of simultaneous action localization and detection was proposed in [5], in which authors used 3D-HOG descriptors on a sliding window. Furthermore, Schiele in [6] used both body pose and motion features for action detection and in [7], web images were used for training a CNN-based activity detector through transfer learning.

While Dense Trajectories (DT) have shown a great potential in action recognition [8, 9, 10], their adoption in Action Detection (AD) remains a challenging task. So far, [1, 11] used DT in a similar manner in this field. In particular, they extracted trajectory features from fixed-length video segments, facing two major issues: first, the splitting is performed uniformly and a significant amount of negative data can be mixed with positive data, and second, finding the optimal length of these splits remains an open challenge, which depends on many parameters, such as speed, action class etc.

In this paper, we propose an effective way to use dense motion trajectories in action detection. Instead of segmenting the video sequences using a sliding window and extracting trajectory features from them, we develop a two-step supervised algorithm for detection and classification. The first step includes the segmentation of the video sequences into temporal regions of interest. This is perfomed by classifying each frame as a positive or negative action. When the action proposals are generated, the second step, which is the classification using improved trajectories, is applied on each generated region. Our contribution is twofold. First, we propose an efficient way of detecting temporal regions of interest in videos which can be coupled with any descriptor for action recognition. Second, we avoid training the classifier with background trajectory data, which usually have low amount of motion and can potentially lead to degradation of performance.

The structure of the paper is the following: the background of our approach is given in Section 2. The proposed model for action detection is described in Section 3. Both the experimental setup and the results are presented in Section 4. Finally, a discussion on results and future steps is presented in Section 5.

This work was funded by the European Union's Horizon 2020 research and innovation project STARR under grant agreement No.689947, and by the National Research Fund (FNR), Luxembourg, under the project C15/IS/10415355/3D-ACT/Björn Ottersten.

2. BACKGROUND

In this section, we briefly review concepts that are used throughout the paper and formulate the problem.

2.1. Improved Dense Trajectories

In order to represent actions in videos, Wang et al. [9] proposed to extract dense motion trajectories for aligning descriptors. This approach starts by uniformly sampling points in the image, and then each point $\mathbf{p}_t = (x_t, y_t)$ at frame t is tracked in the next frame using dense optical flow. A trajectory is defined as a sequence:

$$\mathsf{T}^m_\tau = \{\mathbf{p}^m_{t_0}, ..., \mathbf{p}^m_{t_0+L}\}, \quad (1)$$

where $\tau = [t_0, ..., t_0 + L]$ is the temporal range of the trajectory, $m = 1, ..., M$ is the trajectory index, and L is fixed and set to 15 frames. Trajectories that are static are rejected because they are irrelevant for analysing human actions, while trajectories with large motion are also rejected because they usually correspond to erroneous estimations.

Compared to the original dense trajectories approach [8], authors in [9] propose the removal of camera motion by estimating the homography between consecutive frames using the Random Sample Consensus (RANSAC) algorithm. In this case, Speeded Up Robust Features (SURF) [12] are computed and matched based on the nearest neighbor rule.

In [9], four different descriptors are used for representing videos: the Trajectory Shape Descriptor (TSD) [9], Histograms of Oriented Gradients (HOG), [13], Histogram of Optical Flow (HOF) [13], and Motion Boundary Histogram (MBH) [9]. In order to aggregate the information of the different descriptors and train a classifier for action recognition, Fisher Vectors [14] model is used. Fisher Vectors (FV) encode both first and second order statistics between the video descriptors and a Gaussian Mixture Model (GMM). The individual FV are concatenated and used as input to a Support Vector Machine (SVM) classifier. Since there are multiple action classes to be recognized, a *one-vs.-all* approach is used [9].

2.2. Improved Dense Trajectories of partial actions

The improved dense trajectories with Fisher Vectors (iDT+FV) approach has shown great potential in action recognition thanks to two major advantages. Initially, the dense optical flow field offers a low-level motion analysis for videos without additional cost. Secondly, the tracking of fast and irregular motion patterns is robust since the optical field is being smoothed.

However, iDT+FV works inadequately when only a partial view of an action is available (refer to Table 1). In this case, the motion pattern is not descriptive and can lead to incorrect classification. The previous observation makes the use of iDT+FV in action detection particularly challenging. According to a study we conducted on iDT+FV on partial actions, the utilization of fixed window approach seems to be insufficient. In particular, we applied two classification approaches on the MSR DailyActivity 3D dataset [15]. In Video Segmentation approach, videos are divided into segments of equal length and from each segment we extract the iDT+FV features. During classification, we utilize one SVM per video segment. In the Features Grouping approach, we first extract the iDT+FV features from the video sequences and then we group them together, following the above classification method. The mean accuracy results of both cases are shown in Table 1. As a reference point, we used the mean accuracy measure from the standard iDT+FV approach. The obtained results suggest that partial action recognition using the iDT+FV approach is a particularly challenging task which leads us to the conclusion that this approach is not suitable for action detection.

Table 1. Mean accuracy results on action recognition using Video Segmentation and Features Grouping approaches.

Segments/video	Video Segm.	Features Group.
1 (baseline)	63.75%	
2	63.12%	60.62%
3	58.75%	65.00%
4	61.25%	60.62%

3. PROPOSED MODEL

Our goal is to create a trajectory-based action detection model which addresses the challenges discussed in Section 2.2. This is accomplished using a two-step supervised model: During the first step, a frame-based binary classifier extracts the action proposals from the video sequences and, during the second step, these proposals are classified using a second classifier trained on trajectories features. The idea is to perform a non-uniform video segmentation which can detect full-length video proposals instead of partial action views, offering a more appropriate solution for motion pattern descriptors. We propose two approaches in this regard: in the first, 3D skeleton joints are available and the corresponding features are extracted from them, while in the second, we assume that only RGB video sequences are given, therefore a heatmap of likelihood scores of skeleton joint locations is used. Regarding action classification, a second classifier is employed, trained on full-length action clips. The general pipeline of our approach is given in Fig. 1.

3.1. Video Segmentation using 3D features

The first method makes use of the explicit 3D skeleton joints $\mathbf{s}^j_t = (\mathbf{x}^j_t, \mathbf{y}^j_t, \mathbf{z}^j_t)$, where j is the index of a particular joint

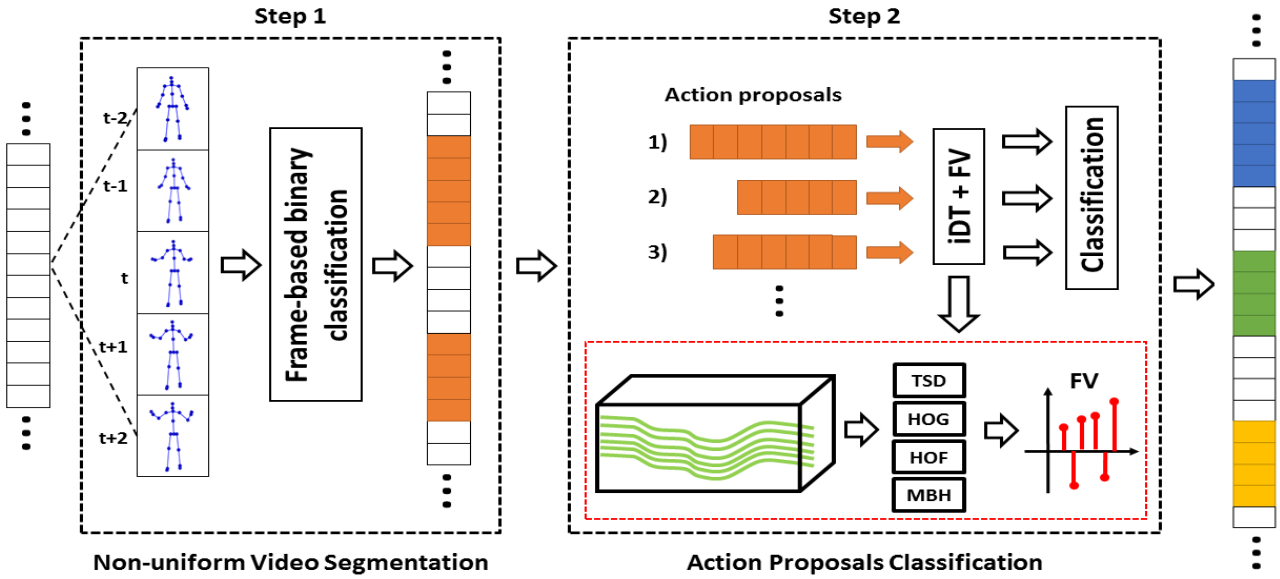

Fig. 1. Our proposed model for action detection. During Step 1, we extract skeleton joint features (or likelihood areas of joints in 2D case) from a temporal window around the current frame and use them as input to a classifier in order to generate the action proposals from the input sequence. During Step 2, standard action recognition using improved trajectories is performed on the action proposals, resulting in the final labeled sequence.

such that $j \in [1, ..., J]$ and t is the current video frame. In order to derscribe 3D poses, we use relative joint position features, as appeared in [16]. Those features are generated by computing the distance between each pair of 3D joints:

$$\delta_t^{ij} = \mathbf{s}_t^i - \mathbf{s}_t^j, \quad \forall\, i, j \in [1, ..., J]. \tag{2}$$

The second descriptor used in this context is similar to the Histogram of Oriented Displacements (HOD) [17]. The HOD concept describes the orientation of each 3D skeleton joint as three 2D trajectories, one for each orthogonal Cartesian plane (xy, xz, yz). For each Cartesian plane, a direction angle θ_j is computed along a temporal window w, as shown below:

$$\theta_j = tan^{-1}\left(\frac{d(j_{xz})}{d(j_{xy})}\right), \tag{3}$$

where $d(j_{xz})$ and $d(j_{xy})$ are the spatial distances of joint j between consecutive frames in Cartesian planes xz and xy correspondingly. The orientation features are computed between consecutive frames for a temporal window w round the current frame. The histogram representation is the accumulation of the motion orientation in the quantitized 2D space.

Finally, a binary k-Nearest Neighbor (kNN) classifier is employed for labeling each frame. This classifier seems to be the most suitable solution for our concept, because of its balance between simplicity and high accuracy.

3.2. Video segmentation using 2D features

In this approach, we assume that the 2D pose information is not provided and only RGB video sequences are available. Therefore, a state-of-the-art human pose detector [18] is used for the estimation of the likelihood areas of the 2D body joints. This Convolutional Neural Network (CNN)-based pose detector provides a likelihood heatmap C_j of the joint position j at frame t. These heatmaps are concatenated and used as pose features. They, also, seem to be more tolerant to erroneous estimation of body pose than raw 2D joints [19].

In addition, the computation of motion features in this context relies on the HOD descriptor, as shown in (3). In order to employ it, we need to estimate the exact position of body joints. Therefore, we max-pool the likelihood scores from every joint heatmap, as shown below:

$$s^j = argmax(C_j). \tag{4}$$

In this case, however, only the xy Cartesian plane is used for describing the motion evolution. Similarly to Section 3.2, a binary kNN classifier was utilized for frame-based labeling.

3.3. Action proposals classification

As a pre-classification step, we correct the continuity of action proposals. We apply median filter to classifier's scores and re-generate the labels. In addition, window-based patching of labeled sequences is applied for filling any blank areas within the detected action proposals.

The second step, as shown in Fig. 1, is common for both approaches. During this step, we apply the standard iDT+FV approach for the classification of action proposals. The motion trajectories descriptors are the same used in [9]; TSD, HOG, HOF and MBH. Moreover, a *one-vs-all* linear SVM classifier is trained on groundtruth action clips.

4. EXPERIMENTS

Improved trajectories and motion descriptors are computed using the implementation provided in [9][1]. In addition, in Section 3.2 the pre-trained CNN-based human pose detector [18][2] was used for obtaining the 2D body pose heatmap.

For generating action proposals, we propose two different approaches. The first proposed approach is called **Skeleton-based Segmentation** and uses the 3D skeleton joint descriptors for action proposals generation, as described in Section 3.1. Our second proposed approach, called **Heatmap-based Segmentation** and presented in Section 3.2, utilizes the likelihood scores C_j provided by the output of the CNN-based pose estimator in order to compute the 2D features for action proposals generation.

Our approaches are evaluated on the Online Action Detection dataset [20]. It consists of 10 daily action classes *(drinking, eating, writing, opening cupboard, opening oven, washing hands, sweeping, gargling, throwing trash* and *wiping*) captured continuously and mixed with a large amount of background motion. The sequences were captured using a Kinect v2 sensor, thus RGB, depth and 3D skeleton joint data are available. We follow the dataset's splitting protocol into training and testing sets.

In both approaches, we use 11 frames window length for pose descriptors and 21 frames window length for the motion descriptors. These numbers were obtained by cross-validation. Moreover, we used 8 bins for computing the HOD features. The 3D body pose (used in the proposed **Skeleton-based Segmentation** approach) of the Online Action Detection dataset consists of 25 joints, whereas in the **Heatmap-based Segmentation** case, the 2D body pose is described by 16 likelihood areas of joints.

For measuring the performance of our approaches, we used the F1-score measure, which is defined as:

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} \quad (5)$$

We used as baseline the iDT+SW [11] approach. The obtained results for this model on the Online Action Detection dataset are shown in detail in Table 2. The average F1-score is 0.467 which is lower than the proposed **Heatmap-based Segmentation** approach by 0.076. The superior performance of our model is justified by the fact that it addresses the two major issues of iDT-based approaches: full action proposals are provided as input to the classifier which are more discriminative compared to partial action segments and most of the background (negative) frames are removed from the video segments. Using the more informative 3D joint descriptors in the **Skeleton-based Segmentation** case, we achieved a significant performance improvement. In particular, we reached an average F1-score of 0.671, which is higher than the JCR-RNN's reported performance in [20] (0.653 - refer to Table 2). Despite the fact that the classification was performed on 2D data (RGB frames), the action proposals were significantly more accurate (88.28% frame-based detection accuracy) and compensated the absence of 3D data in Step 2.

Table 2. F1-score results for Heatmap-based and Skeleton-based approaches against JCR-RNN and iDT+SW on Online Action Detection Dataset.

	JCR-RNN [20]	**iDT+SW** [11]	**Heatmap-based**	**Skeleton-based**
Drinking	0.574	0.350	0.218	0.568
Eating	0.523	0.353	0.404	0.484
Writing	0.822	0.582	0.619	0.792
Opening cupboard	0.495	0.453	0.499	0.669
Opening oven	0.718	0.294	0.581	0.677
Washing hands	0.703	0.591	0.759	0.714
Sweeping	0.643	0.467	0.430	0.800
Gargling	0.623	0.505	0.550	0.619
Throwing trash	0.459	0.425	0.573	0.548
Wiping	0.780	0.647	0.802	0.842
Average	**0.653**	**0.467**	**0.543**	**0.671**

5. CONCLUSION

In this paper, we proposed a novel procedure to use improved trajectories for action detection, by pre-defining the temporal regions of interest. The improved performance comes from Step 1, where the generation of action proposals along with the removal of background frames take place. The positive impact of Step 1 is noticeable in iDT+FV classification of Step 2. The recognition of action proposals becomes more precise, since negative data are widely removed and actions are fully visible. The obtained results (Table 2) show our model's superiority over some noteworthy approaches in action detection. As future work, we intend to extend the current approach to 3D motion trajectories and empower the second step of our model by adding viewpoint invariance to it.

[1] https://lear.inrialpes.fr/people/wang/improved_trajectories

[2] https://fling.seas.upenn.edu/~xiaowz/dynamic/wordpress/monocap/

6. REFERENCES

[1] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars, "Online action detection," in *European Conference on Computer Vision (ECCV)*, 2016.

[2] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre, "Sequential max-margin event detectors," in *European Conference on Computer Vision (ECCV)*, 2014.

[3] Minh Hoai and Fernando Torre, "Max-margin early event detectors," in *International Journal of Computer Vision (IJCV)*, 2014.

[4] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid, "Actom Sequence Models for Efficient Action Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011.

[5] Alexander Kläser, Marcin Marszalek, Cordelia Schmid, and Andrew Zisserman, "Human focused action localization in video," in *European Conference on Computer Vision (ECCV)*, 2012.

[6] Bernt Schiele, "A database for fine grained activity detection of cooking activities," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from web images," in *ACM Multimedia Conference (MM)*, 2015.

[8] Heng Wang, A. Kläser, C. Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[9] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[10] Limin Wang, Yu Qiao, and Xiaoou Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] Zhixin Shu, Kiwon Yun, and Dimitris Samaras, "Action detection with improved dense trajectories and sliding window," in *European Conference on Computer Vision Workshop (ECCVW)*, 2015.

[12] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, June 2008.

[13] Cordelia Schmid, Benjamin Rozenfeld, Marcin Marszalek, and Ivan Laptev, "Learning realistic human actions from movies," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[14] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision (ECCV)*, 2010.

[15] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human Action Recognition by Representing 3D Human Skeletons as Points in a Lie Group," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] Mohammad A. Gowayyed, Marwan Torki, Mohamed E. Hussein, and Motaz El-Saban, "Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition," in *International Joint Conferences on Artificial Intelligence*, 2013.

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," *European Conference on Computer Vision (ECCV)*, 2016.

[19] Konstantinos Papadopoulos, Michel Antunes, Djamila Aouada, and Björn Ottersten, "Enhanced trajectory-based action recognition using human pose," in *IEEE International Conference on Image Processing (ICIP)*, 2017.

[20] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *European Conference on Computer Vision (ECCV)*, 2016.