UNSUPERVISED ACTION SEGMENTATION OF UNTRIMMED EGOCENTRIC VIDEOS

Sam Perochon, Laurent Oudre

Université Paris Saclay, Université Paris Cité, ENS Paris Saclay,
CNRS, SSA, INSERM, Centre Borelli, F-91190, Gif-sur-Yvette, France

ABSTRACT

The introduction of affordable wearable cameras and eye trackers have led to a massive amount of egocentric (or first-person view) videos, bringing new challenges to the computer vision community for understanding and leveraging the specificities of the egocentric view. This work proposes a novel approach for unsupervised activity segmentation that detects frames corrupted by ego-motion and estimates action boundaries using kernel change-point detection. The approach leverages the visual characteristics of egocentric videos to improve segments' temporal accuracy. We report state-of-the-art performances for unsupervised approaches on two challenging large-scale datasets of untrimmed egocentric videos, EGTEA and EPIC-KITCHEN-55, and on the standard third-person view dataset, 50Salads.

Index Terms— Temporal Activity Segmentation, Temporal Activity Localization, Egocentric Video Understanding, Kernel Change Point Detection

1. INTRODUCTION

Unsupervised action segmentation aims to automatically detect temporal actions boundaries in videos and classify them into action categories (including background segments that do not contain relevant high-level actions), which is needed for various tasks related to the understanding of human activity in videos [1].

Egocentric videos are recorded by cameras, usually mounted in people's heads, body or equipment. They could be used for life recording, educational content creation or clinical applications. They offer rich intrinsic information that is not available in exocentric videos, such as cues of the camera wearer's gaze, cues to head and body movement, occlusion-free interactions with objects, and a clear view of the hand pose [2, 3]. Although these specificities can be exploited for action segmentation, the lack of global context and the presence of ego-motion due to camera movements can potentially degrade the performances of available approaches relying on optical flow extraction [4].

Supervised approaches have received a lot of attention to address temporal segmentation of videos. For instance, transformers-based models [4] and two-stream temporal convolutional networks [5] have been used to model long-range time dependency across the video, potentially post-processing the output segmentation using graph-based models [6]. Weakly-supervised approaches have been proposed to mitigate the cost of annotations and their subjectivity and use either the ordered sequence of actions [7], sparse annotations e.g. labelling only an arbitrary frame within each action segment [8], or human-generated information such as speech or captions [9]. However, these approaches follow a paradigm that relies heavily on training data, which limits their practical value and requires high-quality image-level annotations at the cost of significant, time-consuming and unscalable human effort.

To circumvent this problem, unsupervised approaches have recently emerged. Some work uses alternate self-supervised learning of visual and temporal appearance representations of the frames of multiple videos of the same activity before clustering them to obtain the action segmentation [10]. These two steps can be merged to learn an embedding space while jointly clustering the frames representation using a temporal optimal transport module that preserves detected actions ordering [11]. In the same spirit, actions in multiple videos can be co-localized by iteratively optimizing a clustering of the video frame representation and the localization of action boundaries [12]. However, these works make the restrictive assumption that several videos of the same activity are available. A state-of-the-art approach, TW-FINCH, overcomes this limitation by building a time-weighted nearest-neighbor graph from the frame embedding, and discovers action segments by applying a recursive hierarchical clustering algorithm. Connected components of the graph form clusters with similar appearances [13]. However, it is assumed that the camera is fixed, which might limit their performances on egocentric content.

To the best of our knowledge, our work is the first unsupervised action segmentation method (without any supervision or training) dedicated to egocentric videos. In this work, we hypothesized that in egocentric videos, actions should appear as a succession of frames with similar visual appearance (e.g., containing the same objects). Therefore, we expect features extracted from the images to have high intra-segment similarities and low inter-segment similarities. Conversely, in egocentric videos, the background segments are expected

to be composed of a succession of images with low inter-segment similarities (e.g., due to head or body movements resulting in motion blur in the frames).

The main contributions of this work include a simple yet effective end-to-end approach for unsupervised activity segmentation of long untrimmed videos, blind to action classes and leveraging the specificities of egocentric videos using an ego-motion detection procedure that improves its accuracy.

2. METHOD

Given a video containing N frames, and a vocabulary of $N_{actions}$ actions, we aim to find a set of actions boundaries $\{\theta_1, ..., \theta_P\}$ with $\theta_i = (\theta_i^{start}, \theta_i^{end}) \in [1, N]^2$ and $\theta_i^{start} < \theta_i^{end}$, and their action classes $\{c_1, ..., c_P\} \in [1, N_{actions}]^P$. Note that in the following, P is the *unknown* number of action segments in the video and $N_{actions}$, the *known* number of different actions in the video.

2.1. Feature extraction and Gram matrix computation

The first step of our pipeline consists in extracting features from the N video frames. For this purpose, we use as an off-the-shelf feature extractor the penultimate layer of a pre-trained Resnet152 model [14]. We opted for a Resnet152 model as it was trained on Imagenet, which only contains static images and is therefore expected to build poor representations of blurry frames, which is a desirable property in this work. As a result, the video is transformed into a set of feature vectors $X = \{x_1, ..., x_N\} \in \mathbb{R}^D$, with $D = 1000$.

As stated in the introduction, we aim to perform action segmentation by relying on similarities/dissimilarities of the frames. To this end, we construct the Gram matrix defined as

$$K_{i,j} = \frac{x_i^T x_j}{\sqrt{\|x_i\| \|x_j\|}} \quad (1)$$

where the cosine similarity coefficients are comprised between -1 and +1. As illustrated in Figure 1, this matrix provides important essential for the temporal segmentation of the video. Diagonal blocks of high similarities encode segments of frames with consistent visual appearances, expected to be associated with atomic actions. Extra-diagonal blocks of high similarities encode redundant actions happening before of after in the video. Finally, diagonal blocks with low similarity values are associated with successive frames with different appearances, which is hypothesized to be caused by ego-motion. The latter are associated with background segments as they usually do not contain actions.

2.2. Preliminary detection of background frames

The Gram matrix K can be used to detect frames with a low local similarities, which will be labelled as belonging to the background class. A frame i is said to have low local similarity if

$$\max \left\{ \underset{|j-(i-\Delta)|<\delta}{\text{mean}} K_{i,j}, \underset{|j-(i+\Delta)|<\delta}{\text{mean}} K_{i,j} \right\} < \alpha, \quad (2)$$

with $\alpha = 0.75$, $\Delta = 0.5$ sec, and $\delta = 0.1$ sec. This step results in a binary mask with 1 for background frames and 0 otherwise. A final smoothing of this binary mask is applied, such that isolated consecutive background frames with lengths less than 0.15 sec are removed from the set of background frames, while gaps of size less than 0.3 sec surrounded by background frames are labelled as background.

2.3. Temporal action boundaries estimation through kernel change-point detection

In order to segment the video, we rely here on a change-point detection method [15] using a kernel cost function [16].

Let us consider that the feature vectors x_i were mapped onto a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with the cosine similarity kernel $k(.,.) : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. The function $\phi : \mathbb{R}^D \to \mathcal{H}$ that maps a feature vector to its embedding in the RKHS is implicitly defined by $\phi(x) = k(x, .) \in \mathcal{H}$, resulting in the following inner-product and norm: $< \phi(x_i)|\phi(x_j) >_{\mathcal{H}} = k(x_i, x_j)$ and $\|\phi(x_i)\|^2_{\mathcal{H}} = k(x_i, x_i)$.

The approach consists in finding mean-shifts in the mapped signal $(\phi(x_1), ..., \phi(x_N))$ by minimizing the following objective function:

$$V(t_1, ..., t_P) = \sum_{p=0}^{P} \sum_{t=t_p}^{t_{p+1}-1} \|\phi(x_t) - \bar{\mu}_{t_p,...,t_{p+1}}\|^2_{\mathcal{H}}, \quad (3)$$

where $\bar{\mu}_{t_p,...,t_{p+1}}$ is the empirical mean of the segment $(\phi(x_{t_p}), ..., \phi(x_{t_{p+1}})$, and $t_1, ..., t_P$ the ordered change point indexes. Note that thanks to the kernel tricks, all terms in (3) can be expressed with the matrix $K_{i,j}$

Since the number of change points is not known beforehand, a regularization term is added to form the penalized optimization problem

$$\hat{P}, \{\hat{t}_1, ..., \hat{t}_P\} = \text{argmin}_{P,t_1,...,t_p} V(t_1, ..., t_p) + \lambda P, \quad (4)$$

with $\lambda > 0$ the smoothing parameter and $\hat{P}$ the number of estimated change points. A fast and efficient resolution of this optimization problem is obtained in $\mathcal{O}(N)$ using dynamic programming and the PELT method [17].

Note that frames belonging to the background segments (identified in Step 2) are withdrawn before performing the change-point detection procedure to avoid the presence of outliers in the data that could false the detection. After the segmentation procedure, the background segments boundary indexes are added to the set of estimated change points.

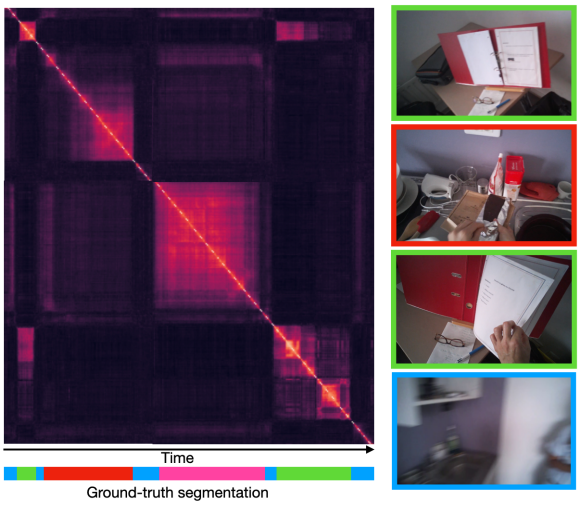

Fig. 1. Sample Gram matrix, ground-truth (GT) segmentation, and sample frames.

2.4. Clustering of the segments

We use a clustering procedure to assign a class value to each segment. First, the feature vectors of dimensions $D = 1000$ are reduced to 100 dimensions using a Principal Component Analysis (PCA). We then use the standard K-means algorithm to assign each frame to one of the $N_{actions}$ activity clusters. Finally, each segment is associated with the cluster index that is the most represented within the segment.

Note that, similarly to the segmentation step, we remove the background frames before performing the clustering procedure, thus avoiding the presence of outliers.

2.5. Final detection of background segments

After this process, some background frames may still be present in the retrieved segments. To remove these spurious frames from the set of actions, we use a byproduct provided by the clustering procedure, i.e. the silhouette score of each segment [18]. For each feature vector x_i we compute a silhouette score defined as

$$S(x_i) = \frac{s_{C_i}(x_i) - s_{\tilde{C}_i}(x_i)}{\max\{s_{C_i}(x_i), s_{\tilde{C}_i}(x_i)\}}, \quad (5)$$

with $s_C(x_i) = \frac{1}{|C|}\sum_{j \in C} K_{ij}$ being the mean similarity between x_i and the other x_j from cluster C, and C_i and $\tilde{C}_i$ being respectively the cluster assigned to x_i and the next closest cluster to x_i. If the average silhouette score within an action segment is lower than a standard deviation below the average silhouette value, the segment is labelled as a background segment. In practice, we found that this segment's post-processing step removes segments corresponding to transitions between activities, hence being poorly associated with well-defined activity clusters.

3. EXPERIMENTAL SETTINGS

3.1. Datasets

We tested our approach on two large-scale datasets of egocentric videos EGTEA [19] and EPIC-KITCHEN-55 [20], which are considered challenging because they contain videos longer than 10 minutes with over 100 actions per video. The EGTEA dataset consists in 86 videos and has a total duration of 29 hours, with a large proportion of background frames ((45%). The EPIC-KITCHEN-55 data set consists of 55 hours of unscripted activities of daily living. Since the ground-truth of the test set is not available, we followed standard practice and report the results on participants 26 to 37 of the training set (130 videos; with 27% of background frames) [6]. In addition to these two datasets, we provide results for the exocentric dataset 50Salads [21], which consists in 50 videos in third-person view of 25 different actors performing a cooking activity, with a total duration of approximately 4.5 hours and 14% of background frames.

3.2. Performance metrics.

Following standard unsupervised approaches [13], we used the Hungarian matching algorithm to match the predicted actions cluster to the ground truths activities [22]. We evaluated the performance of our approach using standard segment-wise metrics, namely the $f1@\{10, 25, 50\}$ score. For this metric, a predicted segment is labelled as correct only if the action class is correct and if the overlap with a potential ground truth action is higher than $k\%$, with the overlap measured with the intersection over the union. We also report standard frame-wise metrics, namely the accuracy measured as the mean over frames (MoF), and the edit score.

3.3. Implementation details.

To be consistent with the other benchmark methods, we evaluated our approach on the verbs, actions, and eval classes of EGTEA, EPIC-KITCHEN-55, and 50Salads, respectively, and assumed that the number of actions $N_{actions}$ in the videos was known [9]. We compared our approach with the two recent supervised (m-GRU+GTRM; [6] and unsupervised (TW-FINCH; [13]) SoTA approaches. We used the author's source code of TW-FINCH, without change of parameters and used the Resnet 152 feature vectors to evaluate their performances. The video duration differences between the three datasets require different values of the penalty parameter λ in the equation (4), which is proportional to the number of video frames, with a value $\lambda = 10$ for feature vectors of size $N = 3000$. The source code is available and can readily be applied to long untrimmed videos.

4. RESULTS AND DISCUSSION

Table 4 shows the performance of our approach on the benchmark datasets. The supervised approach performed better on **EGTEA**, with for instance, an $f1@10$ score of 41.6 against 24.3 for method, and 12.9 for TW-FINCH. However, we outperformed TW-FINCH by achieving $f1$ scores 2 to 6 times higher when tightening the overlap constraints, with a $f1@50$ gain of 10 ($2.4 \rightarrow 12.4$)), demonstrating its advantages over SoTA unsupervised approaches. As for the **EPIC-KITCHEN-55** dataset, both unsupervised approaches outperformed the supervised one, with e.g. a $f1@50$ gain of 10.7 ($10.7 \rightarrow 21.4$)) for our method, showing that the use of labelled training data is not always necessary. However, the performance between the unsupervised approaches need to be qualified as our approach only reached the best performances in MoF (43.7% against 42.8% for TW-FINCH) and in the most constrained $f1@50$, while the $f1@10$ and $f1@25$ are more advantageous for TW-FINCH ($42.0 \rightarrow 51.7$ and $35.7 \rightarrow 40.5$, respectively). The difference is mainly due to the fact that this dataset, compared to EGTEA, contains more short actions (e.g. the median duration of actions are 1.45s and 2.1s, respectively). However, as illustrated in Figure 2, TW-FINCH tends to output actions segment with little variation in duration, thus avoiding potential false detections of short actions, unlike our approach, which tends to detect actions of any length and, therefore, may suffer from a larger number of false positives. This effect decreases when the overlap constrain is 50% as TW-FINCH, therefore, has less true positive and more false negative. As can be seen, our approach outperforms TW-FINCH on **50Salads** for the f1 measures by 2.8 ($37.7 \rightarrow 40.5$) 3.2 ($34.0 \rightarrow 37.2$) and 1.4 ($23.1 \rightarrow 24.5$) for the $f1@10, 20$, and 50, respectively. This demonstrates that our approach can be used for third-person view videos which were not the main focus of this work. Notably, this shows the robustness of the background detection steps of our approach, which did not detect a lot of background segments in a dataset that contains only 14% of them.

For the datasets with egocentric videos, which is the focus of this work, the performances improvement compared to the SoTA approach TW-FINCH is correlated with the percentage of background frames, which highlights the efficiency of our background detection steps. Indeed, an ablation study on the **EGTEA** dataset suggested that the averaged segmental $f1$ measure improved by 8.03 when discarding background segments before the change-point detection step, suggesting its importance both to enhance the kernel-change-point step, and to identify the background segments in a robust way.

Additionally, our approach seems more appealing for applications requiring larger temporal accuracy, as demonstrated by its larger $f1@50$ performance gains on all benchmarked datasets. Interestingly, the performance gap between supervised and unsupervised approaches shrinks or disappears in the case of egocentric datasets, showing that the segmentation of egocentric videos is more challenging and can be addressed with unsupervised method. Finally, our approach has the same advantageous computational complexity than TW-FINCH in $\mathcal{O}(N^2)$ (due to the clustering step).

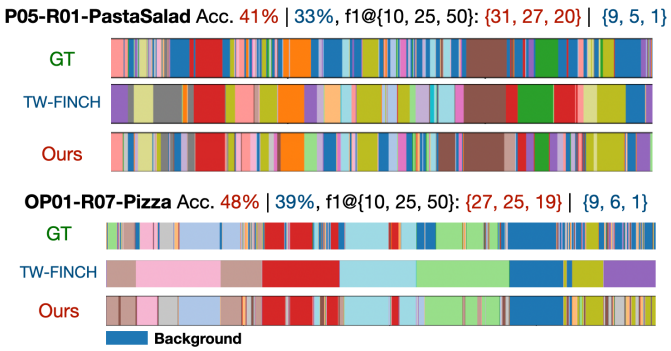

Fig. 2. Cropped segmentation of two videos of the EGTEA [19] dataset, illustrating the ability of our approach to delimit action segments of various duration.

Methods	f1@10	f1@25	f1@50	Edit	MoF
EGTEA					
m-GRU+GTRM[6]	41.6	37.5	25.9	41.8	69.5
TW-FINCH[13]	12.9	6.7	2.4	12.3	33.7
Ours	**24.3**	**19.9**	**12.4**	**26.0**	**36.0**
EPIC-KITCHEN-55					
m-GRU+GTRM[6]	31.9	22.8	10.7	42.1	43.4
TW-FINCH[13]	**51.7**	**40.5**	21.2	**39.8**	42.8
Ours	42.0	35.7	**21.4**	32.4	**43.7**
50Salads					
m-GRU+GTRM[6]	77.4	74.6	65.3	67.8	85
TW-FINCH[13]	37.7	34.0	23.1	**32.6**	**40.4**
Ours	**40.5**	**37.2**	**24.5**	26.1	36.2

Table 1. Performances on the three datasets tested. Light-grey indicates the supervised approach [6], and in grey our and a SoTA unsupervised approach TW-FINCH.

5. CONCLUSION

We presented a fast and efficient method for the unsupervised temporal segmentation of actions in long untrimmed videos. We demonstrated that simple heuristics leveraging ego-motion cues could lead to competitive results for the temporal segmentation of egocentric contents while preserving the approach's interpretability. We obtained promising performances on three benchmark datasets, proving the change-point detection framework's efficiency to propose relevant action segments of different durations, and the benefits of detecting background frames when segmenting egocentric videos. This approach has numerous potential applications, e.g., in clinical settings for assessing the interpersonal behavioral differences when performing standardized tasks.

6. REFERENCES

[1] Elahe Vahdani and Yingli Tian, "Deep learning-based action detection in untrimmed videos: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022.

[2] Yin Li, Zhefan Ye, and James M. Rehg, "Delving into egocentric actions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 287–295.

[3] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras, "Egocentric vision-based action recognition: A survey," *Neurocomputing*, vol. 472, pp. 175–197, 2022.

[4] Chenlin Zhang, Jianxin Wu, and Yin Li, "Actionformer: Localizing moments of actions with transformers," in *European Conference on Computer Vision*, 2022.

[5] Colin S. Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory Hager, "Temporal convolutional networks for action segmentation and detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012, 2017.

[6] Yifei Huang, Yusuke Sugano, and Yoichi Sato, "Improving action segmentation via graph-based temporal reasoning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14021–14031.

[7] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall, "Neuralnetwork-viterbi: A framework for weakly supervised video learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7386–7395, 2018.

[8] Zhe Li, Yazan Abu Farha, and Juergen Gall, "Temporal action segmentation from timestamp supervision," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8361–8370, 2021.

[9] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Rosaura G. VidalMata, Walter J. Scheirer, Anna Kukleva, David Cox, and Hilde Kuehne, "Joint visual-temporal embedding for unsupervised learning of actions in untrimmed sequences," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1237–1246.

[11] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M. Zeeshan Zia, and Quoc-Huy Tran, "Unsupervised action segmentation by joint representation learning and online clustering," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20142–20153, 2022.

[12] Guoqiang Gong, Xinghan Wang, Yadong Mu, and Qi Tian, "Learning temporal co-attention models for unsupervised video action localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] M. Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen, "Temporally-weighted hierarchical clustering for unsupervised action segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11220–11229.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] Charles Truong, Laurent Oudre, and Nicolas Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, pp. 107299, 2020.

[16] Sylvain Arlot, Alain Celisse, and Zaid Harchaoui, "A kernel multiple change-point algorithm via model selection," *Journal of Machine Learning Research*, vol. 20, no. 162, pp. 1–56, 2019.

[17] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, oct 2012.

[18] Peter J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[19] Yin Li, Miao Liu, and James M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 11, pp. 4125–4141, 2021.

[21] Sebastian Stein and Stephen J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA, 2013, UbiComp '13, p. 729–738, Association for Computing Machinery.

[22] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.