Hierarchical Vector Quantization for Unsupervised Action Segmentation

Authors: Federico Spurio[1,3], Emad Bahrami[1,3], Gianpiero Francesca[2], Juergen Gall[1,3]

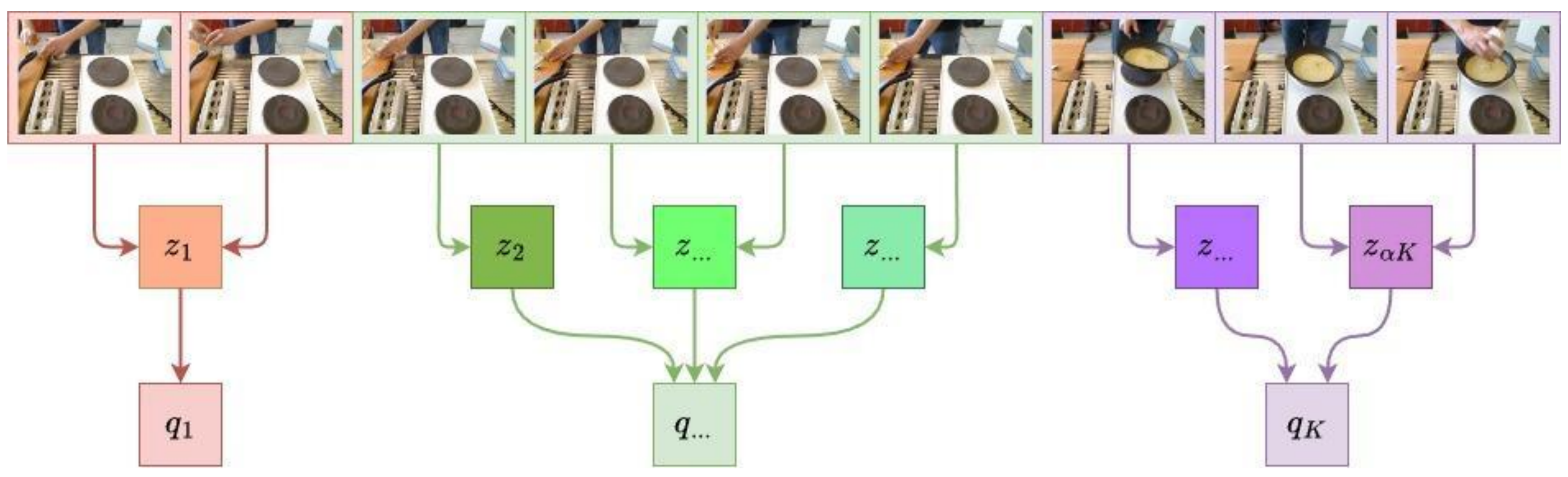
Associated institutes
[1]University of Bonn, [2]Toyota Motor Europe, [3]The Lamarr Institute for Machine Learning and Artificial Intelligence

LAMARR INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Link to our code!

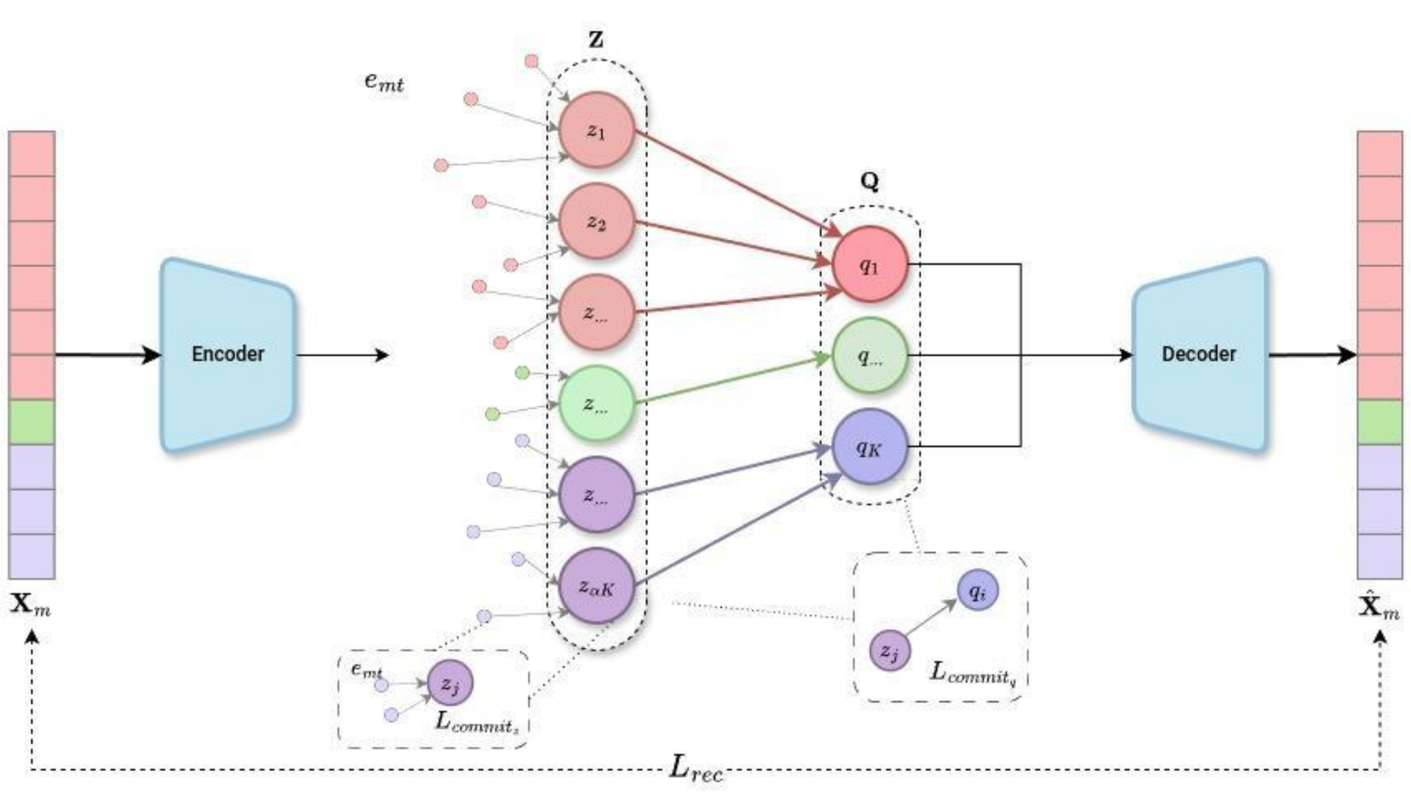
Introduction

- In this work we deal with the task of **unsupervised action segmentation**, which segments a set of long, untrimmed videos into semantically meaningful segments that are consistent across videos.
- Our **key observation**: human actions are of a compositional nature, i.e. intermediate steps are needed to complete a task. This was not taken into consideration in previous works, where they show good performance, at the expenses of distribution of segment lengths.
- To incorporate this idea, we propose to **hierarchically** model action segments using our novel **Hierarchical Vector Quantization** model. Additionally, to measure the quality of distribution of segment lengths compared to the ground truth, we introduce a new metric based on **Jensen-Shannon Divergence**.
- We show that our model achieves **state-of-the-art** results on 3 datasets: Breakfast, YouTube Instructional and IKEA ASM.



Hierarchical Vector Quantization

- We model actions using a **fine-to-coarse hierarchical representation**, capturing both **low-level subactions** and **high-level action structures**.
- Our two-levels quantization maps frames to subaction clusters, a **fine-grained representation** of an action, then grouped into coarse action prototypes to form action representation.
- A **commitment loss** enforces consistency between frames and subations, and between subactions and actions. A **reconstruction loss** ensures meaningful latent representations.
- During inference, each frame is assigned to its **nearest prototype in Q**, and the predictions are refined using **FIFA decoder**.



JSD Metric

- We notice a **bias in terms of the length of the generated action segments** in prior works. For this reason, we introduced **JSD metric**.
- For each video within same activity, we compute the **histogram of the predicted segment lengths**, and compare it to the ground-truth using the **Jensen-Shannon Distance (JSD)**. The JSD scores are **averaged per activity**, then **weighted by the number of frames** to obtain the final score.

Quantitative Results

- We evaluated our approach on 3 datasets: Breakfast, YouTube Instructional (YTI) and IKEA ASM. We achieve **state-of-the-art results** in F1-score, recall and JSD.
- We analyze how the **number of levels of quantization** and the **number of prototypes in Z (α)** affect the predictions.

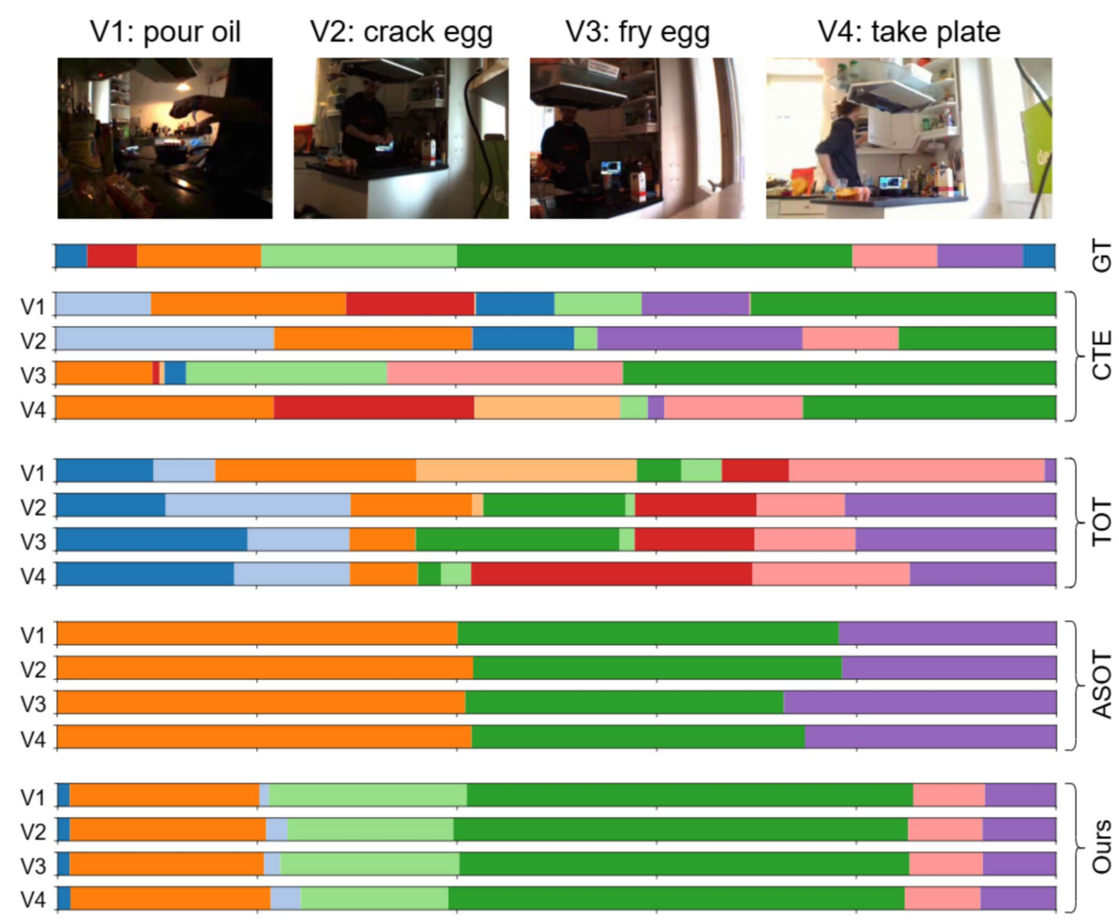
Dataset	Breakfast				YTI			IKEA ASM*			
Method	MOF	F1	Recall*	JSD*↓	MOF	F1	Recall*	MOF*	F1*	Recall*	JSD*↓
CTE	41.8	26.4	27.0	87.4	39.0	28.3	22.1	23.1	22.6	18.9	*73.7*
ASAL	52.5	37.9	-	-	44.9	32.1	-	-	-	-	-
TOT	47.5	31.0	26.3	90.2	40.6	30.0	*31.4*	21.0	20.1	17.1	80.0
TOT+TCL	39.0	30.3	36.0	*85.6*	45.3	*32.9*	27.9	23.8	20.9	17.7	79.5
UFSA	52.1	38.0	-	-	*49.6*	32.4	-	-	-	-	-
ASOT	**56.1**	*38.3*	*40.1*	94.9	**52.9**	**35.1**	27.8	*34.0*	*27.9*	*24.0*	88.7
Ours (HVQ)	*54.4*	**39.7**	**44.9**	**82.5**	50.3	**35.1**	**38.7**	**51.2**	**30.7**	**25.9**	**64.8**

	Breakfast			
	$\alpha=1$	$\alpha=2$	$\alpha=3$	$\alpha=4$
MOF	53.6	**54.4**	52.7	51.8
F1	38.2	**39.7**	38.3	38.2
JSD ↓	83.7	82.5	83.0	**82.2**

Dataset	Metric	Single	Double	Triple
YTI	F1	33.0	**35.1**	31.9
IKEA ASM	F1	25.8	27.6	**30.7**
	JSD	81.9	**62.2**	64.8
Breakfast	F1	37.1	**39.7**	38.2
	JSD	83.1	**82.5**	84.1

Qualitative Results

Segmentation results for a sample of Breakfast. Our approach delivers **highly consistent results** across multiple videos (V1, V2, V3, V4) recorded from different cameras, but with the same ground truth.



References

1. Kuehne, H.; Arslan, A.; and Serre, T. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
2. Alayrac, J.-B.; Bojanowski, P.; Agrawal, N.; Sivic, J.; Laptev, I.; and Lacoste-Julien, S. 2016. Unsupervised Learning From Narrated Instruction Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
3. Ben-Shabat, Y.; Yu, X.; Saleh, F.; Campbell, D.; RodriguezOpazo, C.; Li, H.; and Gould, S. 2021. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).
4. Kukleva, A.; Kuehne, H.; Sener, F.; and Gall, J. 2019. Unsupervised learning of action classes with continuous temporal embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
5. Li, J.; and Todorovic, S. 2021. Action shuffle alternating learning for unsupervised action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
6. Kumar, S.; Haresh, S.; Ahmed, A.; Konin, A.; Zia, M. Z.; and Tran, Q.-H. 2022. Unsupervised Action Segmentation by Joint Representation Learning and Online Clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
7. Tran, Q.-H.; Mehmood, A.; Ahmed, M.; Naufil, M.; Zafar, A.; Konin, A.; and Zia, Z. 2024. Permutation-Aware Activity Segmentation via Unsupervised Frame-To-Segment Alignment. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).
8. Xu, M.; and Gould, S. 2024. Temporally Consistent Unbalanced Optimal Transport for Unsupervised Action Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Partner institutions:

technische universität dortmund
Fraunhofer IAIS
UNIVERSITÄT BONN
Fraunhofer IML

Institutionally funded by:

Bundesministerium für Forschung, Technologie und Raumfahrt
Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen