

# Unsupervised learning of action classes with continuous temporal embedding

## Supplementary Material

Anna Kukleva \*  
University of Bonn  
Germany

s6ankukl@uni-bonn.de

Hilde Kuehne \*  
MIT-IBM Watson Lab  
Cambridge, MA

kuehne@ibm.com

Fadime Sener, Juergen Gall  
University of Bonn  
Germany

sener,gall@iai.uni-bonn.de

| Breakfast dataset   |                  |
|---------------------|------------------|
| Activity class name | # subactions (K) |
| Coffee              | 7                |
| Cereals             | 5                |
| Tea                 | 7                |
| Milk                | 5                |
| Juice               | 8                |
| Sandwich            | 9                |
| Scrambleddegg       | 12               |
| Friedegg            | 9                |
| Salat               | 8                |
| Pancake             | 14               |

Table 1. Number of subactions per complex activity in the Breakfast dataset.

| YouTube Instructions |                  |
|----------------------|------------------|
| Activity class name  | # subactions (K) |
| Changing tire        | 11               |
| Making coffee        | 10               |
| CPR                  | 7                |
| Jump car             | 12               |
| Repot plant          | 8                |

Table 2. Number of subactions per complex activity in the YouTube Instructions dataset.

## 1. Experiments

This document provides additional qualitative results, specifies the number of subactions K, and illustrates the continuous latent space representations.

### 1.1. The number of subactions K

In Tables 1 and 2, we show the numbers of subactions for each activity class. For example, *cereals* activity consists of *SIL*, *take bowl*, *pour cereals*, *pour milk*, *stir cereals* subactions. The 50Salads dataset [2] includes a single complex activity.

\*This work was mainly done at University of Bonn. Asterisk denotes equal contribution.

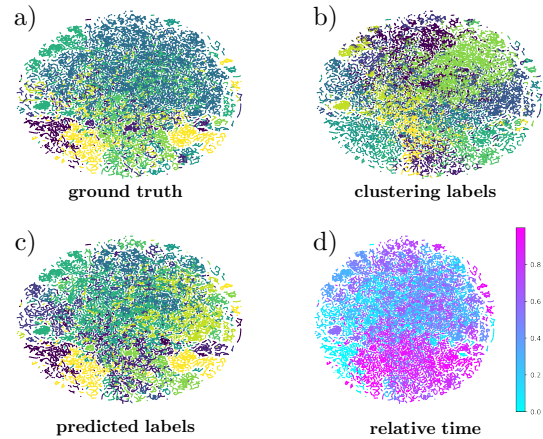


Figure 1. Visualization of embeddings via t-SNE [1] on the 50Salads dataset. Each frame is color coded a) with the corresponding ground truth subaction label, b) with K assigned subaction labels after clustering as the second step in Fig. 1 in our main paper, c) with the predicted labels after the decoding stage. The optimization of our network is performed with respect to relative timestep of each frame. In d) we show the respective relative time label in the continuous temporal embedding assigned to each frame feature. The color bar depicts that dark purple corresponds to 0 (start of the video) and yellow to 1 (end of the video)

### 1.2. Qualitative results

**Latent space** In Fig. 1, we visualize the embedded frame features for the 50Salads dataset. In a), b) and c), colors correspond to the subactions. Subfigure d) shows the encoding of the relative time distribution in the latent space.

**Segmentation** We present additional qualitative results produced by our method in Figs. 2–8. We show some example frames from different subactions, our predicted segmentation for the entire video along with the ground truth segmentations.

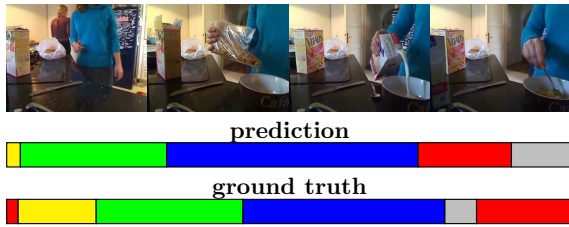


Figure 2. Example for the activity *cereals* of the Breakfast dataset. The order of subactions: *SIL*, *take bowl*, *pour cereals*, *pour milk*, *stir cereals*, *SIL*

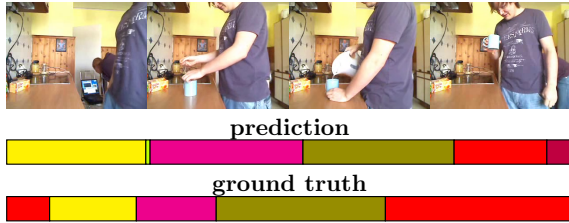


Figure 3. Example for the activity *tea* of the Breakfast dataset. The order of subactions: *SIL*, *take cup*, *add teabag*, *pour water*, *SIL*

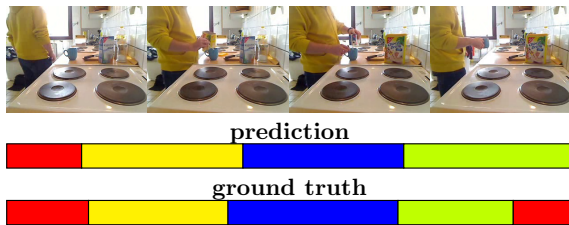


Figure 4. Example for the activity *milk* of the Breakfast dataset. The order of subactions: *SIL*, *spoon powder*, *pour milk*, *stir milk*, *SIL*

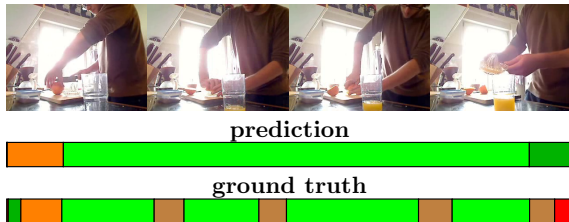


Figure 5. Example for the activity *juice* of the Breakfast dataset. The order of subactions: *SIL*, *take knife*, *cut orange*, *squeeze orange*, *pour juice*, *squeeze orange*, *pour juice*, *squeeze orange*, *pour juice*, *squeeze orange*, *pour juice*, *SIL*

## References

- [1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [2] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food

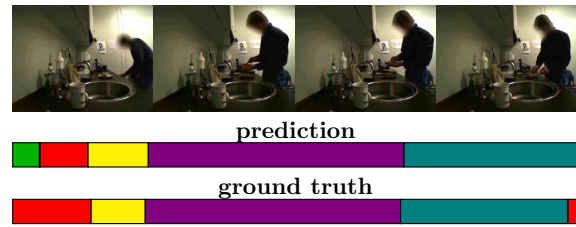


Figure 6. Example for the activity *sandwich* of the Breakfast dataset. The order of subactions: *SIL*, *cut bun*, *smear butter*, *put toppingOnTop*, *SIL*

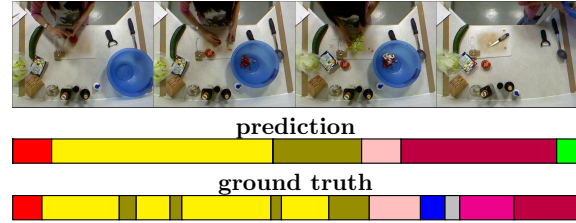


Figure 7. The 50Salads dataset. The order of subactions: *start*, *cut*, *place*, *cut*, *place*, *cut*, *place*, *cut*, *place*, *null*, *null*, *add oil*, *add pepper*, *mix dressing*, *end*

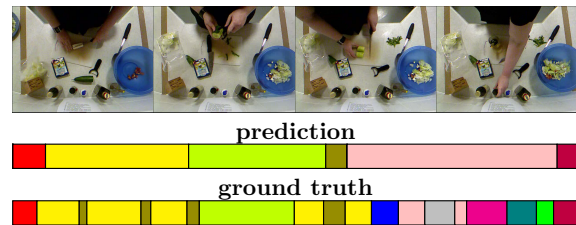


Figure 8. The 50Salads dataset. The order of subactions: *start*, *cut*, *place*, *cut*, *place*, *cut*, *place*, *peel cucumber*, *cut*, *place*, *mix ingredients*, *add oil*, *null*, *add pepper*, *null*, *mix dressing*, *serve salad onto plate*, *add dressing*, *end*

preparation activities. In *UBICOMP*, 2013.