Supplementary materials for OTAS

Anonymous WACV Algorithms Track submission

Paper ID 62

1. Implementation Details

1.1. Global Perception Module

Architecture. The spatial encoder is a ResNet50 model [9] with 2048 output classes. The temporal decoder is adapted from the Transformer [23] with 2048 hidden dimensions and 8 heads. The decoder is constructed by stacking 6 up-sampling blocks. Each block contains an up-sampling function, a convolution layer, and a ReLU activation function. The kernel size is 7 for the first and last blocks and 3 for the rest. The interpolation sizes are 8, 16, 32, 64, 128, and 256, sequentially. The channels are 1024, 512, 256, 128, 64, and 3, respectively.

Optimizer and schedule. We use the standard Adam [10] optimizer with a learning rate of 1e-4 and a multi-step scheduler. We train the model for a total of 250000 steps with a batch size of 16.

1.2. Human-Object Interaction Model

Architecture. The architecture of the human-object interaction model is the same as the global perception model without the decoder.

Human-object interaction masks. We obtain the masks from an off-the-shelf object detection model implemented by the open source platform Detectron2 [28]. We select the Faster-RCNN-X101-FPN model pre-trained on the COCO train2017 dataset [17] with a box average precision (box AP) of 43.0. For human-object interaction, we select masks that contain human body parts, e.g., person.

1.3. Object Relationship Model

Architecture. The encoder is also a ResNet50 model [9], and the decoder is the same as the frame prediction model. We adapt architecture from GATv2 [3] for graph implementation in the bottleneck part. We use two 8-heads self-attention layers, with 32 input channels and 6 output channels for each head. We then add a fully-connected layer to project the output to 2048 dimension.

Optimizer and schedule. We use the standard Adam [10] optimizer with a learning rate of 5e-5 and a multi-step scheduler. We train the model for a total of 100000 steps with a batch size of 16.

Object relationship look-up table. For all the object classes in COCO dataset [17], we select 44 classes that appear most frequently in instructional videos. All 44 classes are depicted in Table 1. We then seek their relations through human annotations from the Visual Genome dataset [11], which is designed for cognitive tasks. To be more specific, we first re-organize the object classes of the Visual Genome dataset to be in line with the 44 classes we selected from the COCO dataset. Some examples of the re-organization are shown in Table 2. Then, for each class, we list and count all possible connections of the objects through predicates provided by the Visual Genome dataset. Finally, we filter out object pairs that appear less than 30 times and build the object relation look-up table. A few illustrations of the look-up table are illustrated in Table 3.

<table>
<thead>
<tr>
<th>toothbrush</th>
<th>scissors</th>
<th>vase</th>
<th>clock</th>
<th>book</th>
<th>refrigerator</th>
<th>sink</th>
<th>toaster</th>
<th>oven</th>
<th>microwave</th>
<th>cell</th>
<th>keyboard</th>
<th>remote</th>
<th>mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>laptop</td>
<td>tv</td>
<td>table</td>
<td>plant</td>
<td>couch</td>
<td>chair</td>
<td>cake</td>
<td>donut</td>
<td>pizza</td>
<td>hotdog</td>
<td>carrot</td>
<td>broccoli</td>
<td>orange</td>
<td>sandwich</td>
</tr>
<tr>
<td>apple</td>
<td>banana</td>
<td>bowl</td>
<td>spoon</td>
<td>knife</td>
<td>fork</td>
<td>cup</td>
<td>glass</td>
<td>bottle</td>
<td>suit</td>
<td>handbag</td>
<td>backpack</td>
<td>bench</td>
<td>person</td>
</tr>
</tbody>
</table>

Table 1. New object classes selected from the COCO dataset.

Object masks. We obtain the object masks from the same off-the-shelf object detection model [28] as the human-object interaction model. We select masks that are in the new object classes and have confidence scores that are larger than 0.7.
2. Evaluation Metrics

2.1. F1 Score

For the computation of the F1 score, we follow the implementation of Shou et al. [20], Wang et al. [25] and first calculate the distance between the \( N \) detected boundaries and the \( M \) ground truth boundaries. We pair each ground truth boundary with a detected boundary that has a minimal distance. Then, we set a fixed distance threshold to determine if the detected boundary is positive or not. The total number of positive detection is \( P \). The Precision/Recall and F1 score can be computed as:

\[
\text{precision} = \frac{P}{N} \\
\text{recall} = \frac{P}{M} \\
F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

We compute the Precision/Recall and F1 score for each video and average across the whole dataset.

As mentioned in the paper, previous works [20, 25] set the distance threshold to be 5% of the length of the corresponding video instance, while we choose 2 seconds which is invariant of video lengths. We show some examples in Figure 1 for the impact of the 2 different thresholds. It is clear that the small threshold is more suitable and general for the evaluation of various instructional videos.

2.2. Hungarian Matching

To perform a fair evaluation with previous methods utilizing clustering algorithms [4, 6, 14, 18], we first applying clustering algorithm as in Du et al. [6] to transfer OTAS boundaries into clusters based on IDT features. Then, we follow [1, 25] and perform the Hungarian matching [13] on a video level.

Noting that for other clustering-based methods that are either only performing on same activities [7, 8, 14, 16, 19, 22, 24, 26, 27] or extend to unknown activities but only provide global-level Hungarian matching results and do not provide code to reproduce [5, 15], we can not conduct a fair comparison.

2.3. Mean over Frames (MoF)

We calculate MoF after clustering and Hungarian Matching. MoF indicates the percentage of frames in the video instance that are correctly segmented [14, 19]. For a video with \( K \) frames, we count all the correct frames \( C \) and compute the MoF as:

\[
\text{MoF} = \frac{C}{K}
\]

We average the video-wise MoF across the whole dataset.

3. User Study

3.1. Implementation

We first pick 20 videos from the Breakfast dataset [12] randomly and generate segmented videos from 5 different methods: one from ground truth, one from OTAS, one from ABD [6], one from CTE [14], and the last one from TW-FINCH [18]. For each video, we shuffle and label 5 segmentation results with numbers 1-5. We invite 33 users to watch and rank the segmentation results with only the reference to the original videos. A part of the user study questionnaire is depicted in Figure 2. Since it is a temporal segmentation task, the average time of completion is 2.5 hours. We use \( 6 - \text{rank} \) as the score for each method (i.e., rank No.1 has 5 points). A video-wise score distribution is shown in Figure 8. We show one of our segmentation results that gains the highest score in Figure 4.
3.2. Breakfast Ground Truth

We provide more illustrations of the inconsistent ground truth segmentation of the Breakfast dataset [12] in Figure 3.

Figure 3. Inconsistency of the ground-truth. For “Sandwich” activity, the action “Cut bun” can be further segment into “Take bun”, “Take knife”, and “Actually cut the bun”; while the action “Smear butter” can be further segment into “Take butter”, “Take knife”, and “Actually smear the butter”. However, the ground truth annotation provides inconsistent segmentation that sometimes produces larger segments and sometimes smaller segments. Even within a video, the segmentation is inconsistent.

4. Qualitative Result

For a better illustration of boundary evaluation, we assign all different ground truth segments distinct colors within a video regardless of labels.

4.1. Breakfast

We provide more qualitative comparisons with the ground truth of our methods on the Breakfast dataset [12] in Figure 5.

4.2. 50Salads

The qualitative comparison of our methods on both eval-level and mid-level 50salads [21] is illustrated in Figure 6. For eval-level, we compare with baselines ABD [6], CTE [14], TW-FINCH [18], Coseg [25] and groundtruth. For mid-level, we only compare with ABD [6], CTE [14], TW-FINCH [18], and ground truth, since Coseg [25] does not provide mid-level results.

4.3. INRIA

The INRIA dataset [2] is collected from YouTube and segmented with the aid of English transcripts obtained from YouTube’s automatic speech recognition (ASR) system. For all tasks, the ordered sequence of ground truth steps is made by an agreement of 2-3 annotators who have watched the input videos and verified the steps on instruction video websites. Therefore, the rest of the video where no step is assigned is considered background. The percentage of average background frames is 73% of all frames. The background frames are various and complicated, as shown in Figure 7. Since we rely on feature differences for boundary detection, the variation in backgrounds influences the result largely. Moreover, we do not have access to prior knowledge of cluster numbers. Therefore, the result of INRIA is very likely to be over-segmented.

5. Ablation Study

5.1. Comparison of Different $\alpha$

We utilize a hyper-parameter $\alpha$ to control the number of boundaries. The comparison of different $\alpha$ is shown in Tab. 4. Generally, lower $\alpha$ leads to higher recall, but also redundancy, which causes precision to drop. Higher $\alpha$ generates fewer boundaries, resulting in higher precision and MoF but low recall. We select $\alpha = 15$ that best balances the trade-off.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>F1(small)</th>
<th>Recall(small)</th>
<th>Precision(small)</th>
<th>MoF</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>42.43</td>
<td>71.96</td>
<td>30.08</td>
<td>65.22</td>
</tr>
<tr>
<td>25</td>
<td>42.77</td>
<td>48.31</td>
<td>38.37</td>
<td>67.57</td>
</tr>
<tr>
<td>15</td>
<td>44.49</td>
<td>53.90</td>
<td>37.87</td>
<td>67.90</td>
</tr>
</tbody>
</table>

Table 4. Comparison of different $\alpha$s. There is a trade-off between better precision and better recall.

5.2. Global Perception Module Architectures

We also conduct an ablation study on different model architectures for the global perception module. Specifically, we leverage features from pre-computed IDT, a pre-trained ResNet-50 model and ResNet with a 2-layer LSTM model that respectively replaces the Transformer layer for comparison. The results demonstrated in Table 5 indicate that the Transformer-based model generates finer features for action segmentation than the other models.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>F1(small)</th>
<th>Recall(small)</th>
<th>Precision(small)</th>
<th>MoF</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>42.43</td>
<td>71.96</td>
<td>30.08</td>
<td>65.22</td>
</tr>
<tr>
<td>25</td>
<td>42.77</td>
<td>48.31</td>
<td>38.37</td>
<td>67.57</td>
</tr>
<tr>
<td>15</td>
<td>44.49</td>
<td>53.90</td>
<td>37.87</td>
<td>67.90</td>
</tr>
</tbody>
</table>

Table 5. Comparison of different $\alpha$s. There is a trade-off between better precision and better recall.
Figure 4. **More consistent granularity of the segmentation results produced by OTAS.** The video shown in the figure contains several smaller segments at the action level. The ground truth only segments “Take butter” out, and combine the others, which is confusing while watching. Furthermore, the ground truth does not separate “Take ingredients” and “Serve on plate” from the backgrounds. However, our segmentation result is neat and consistent, which is more in line with human consensus.

Figure 5. **Qualitative comparison with ground truth (GT) of the Breakfast dataset.** The results predicted by OTAS are largely in line with the ground truth.
Table 5. Ablation of different architectures for the global perception module (OTAS excluding the local attention module) on Breakfast. Transformer-based approach achieves the best performance of F1 score and IoU.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 (small)</th>
<th>MoF</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDT</td>
<td>27.49</td>
<td>63.50</td>
</tr>
<tr>
<td>Pretrained-ResNet</td>
<td>35.28</td>
<td>65.00</td>
</tr>
<tr>
<td>LSTM</td>
<td>36.00</td>
<td>65.50</td>
</tr>
<tr>
<td>Transformer</td>
<td><strong>37.46</strong></td>
<td><strong>65.99</strong></td>
</tr>
</tbody>
</table>
Figure 6. **Qualitative comparison of the 50Salads dataset.** Note that the original illustration of Coseg is not aligned with the actual timestamp. However, since they do not provide code to reproduce, we roughly resize their illustration for comparison.

Figure 7. **Illustration of the various background frames of INRIA.** It contains frames when the person shows preparation, stops to introduce upcoming steps, illustrates precautions, etc. It also contains shot changes and video editing.
Figure 8. Video-wise user study score.
References


