Weakly-Supervised Video Object Segmentation

Daniel Cosentino
University of Central Florida
danny.cosentino12@knights.ucf.edu

Kevin Duarte
University of Central Florida
kevin_duarte@knights.ucf.edu

Mubarak Shah
University of Central Florida
shah@crcv.ucf.edu

Abstract

Video object segmentation is a standard problem in computer vision. Fully supervised methods rely on training models on pixel-level annotations. The problem is that these annotations are time consuming and costly. This paper performs video object segmentation using only a first frame scribble of an object. We propose a method of generating super-pixels and labeling super-pixels as foreground using the scribble in the first frame. The proposed loss function uses the cosine similarity between super-pixels to identify the rest of segmentation. We found in preliminary results that our model achieved an overall score of 0.16 on the YoutubeVOS data set compared to a baseline of 0.31.

1. Introduction

Video segmentation has many applications in video understanding and processing. Video object segmentation (VOS) involves segmenting an object throughout a video given only the first frame segmentation of the object. Segmenting an object throughout a video allows for object tracking and ease of video editing. Fully supervised methods involve using pixel level annotations to train. Pixel level annotations, however, are expensive and time consuming to train. An alternative is to use simpler annotations. Scribbles have been used in image segmentation [4], but have only recently been applied to video [2][5]. However these methods currently involve live interaction with the video, gradually refining the segmentation by drawing scribbles.

Scribbles are used as annotations in image segmentation [4]. The authors used super-pixels along with the scribble to identify the segmentation mask of an object. Scribbles have been used in video object segmentation but only interactively. In [5], an user provides a scribble for the first frame then can repeatedly provide additional feedback, such as for scribbles on false-positives and false negatives, to refine the segmentation masks. This involves constant interaction and for an user to supervise the segmentation masks.

The goal of this research is given just a scribble of an object in the first frame, segment the entire object throughout the video. There is no user feedback and no interaction with the user. A convolutional LSTM network is used based on [8] as the base network. Our method takes in a scribble and generates an initial segmentation based on the scribble and generated super-pixels of the frame. Our proposed loss function then identifies super-pixels that have a high similarity with the foreground super-pixels to fully segment the entire object. This research proposes and demonstrates a loss function that can be applied to weakly-supervised video object segmentation.

2. Weakly-Supervised Learning

2.1. Scribbles

Scribbles are a simpler annotation that consists of a line or segment drawn through the object. They benefit by being much quicker to annotate an object because you are not annotating each pixel in an object. Scribbles are generated for the YoutubeVOS data set for each segmentation by first applying a 12x12 linear convolution over each segmentation. Five random points are then selected in the segment and connected to form the scribble. Scribbles have a thickness of seven pixels. The convolution is applied to prevent from selecting any edge pixels of the annotation.
2.2. Super-pixels

Super-pixels are a way to capture and group image features. SLIC [1] is a state of the art method for generating super-pixels with few parameters and with low computational time. The only parameters are compactness and number of super-pixels. For this research, 200 super-pixels are generated for each frame with a compactness of 22. Generated super-pixels are then labeled as foreground if a scribble goes through the super-pixel and background otherwise. Figure 2 displays this process and the outputted annotation. It is important to note that this method does not fully segment the entire object.

2.3. Objective Function

The proposed loss function is shown in equation 1. The loss is a cross entropy loss with an added alpha term and added regularization.

\[ -\sum Y_i \log(\hat{Y}_i) + (1 - Y_i)(1 - \hat{Y}_i)\alpha + \lambda \| M_{ij} \| \] (1)

Where:

\[ \alpha = 1 - M_{ij}^* \] (2)

\[ M_{ij} = \frac{Y_i \ast Y_j}{\| Y_i \| \ast \| Y_j \|} \] (3)

\[ j^* = \arg\min(M_{ij})|Y_j = 1 \] (4)

Where \( Y_i \) and \( \hat{Y}_i \) are the generated and predicted super-pixels indexed at i. The generated annotations using scribbles do not fully annotate the object. The added alpha term prevents those super-pixels that were labeled as background, but are foreground, to have a loss close to zero. These super-pixels will have a high similarity with the foreground super-pixels and so the alpha term will be close to zero.

The added regularization term is needed to prevent the network from labeling everything as foreground. In this situation the alpha term will be zero for all super-pixels and the total loss will be zero. The regularization term ensures that the model learns the given object segmentation correctly.

2.4. Network Architecture

The network used in this research is the YoutubeVOS Sequence to Sequence model [8]. This network is a sequence to sequence network that uses a convolutional LSTM to segment the object in future frames. First, the network concatenates and encodes the first frame and segmentation to get the initial states of the ConvLSTM. Each subsequent frame is encoded and passed into the ConvLSTM. Both the initializer and the encoder are a VGG-16 [7] network without the fully connected layers. Two convolutional layers and one convolutional layer are appended to the VGG-16 respectively for the initializer and the encoder. The output is then passed through five up sampling layers to produce the predicted segmentation mask. Refer to [8] for further network details.

3. Experiments

3.1. YoutubeVOS

YoutubeVOS [9] consists of 4,453 videos, with 197,272 annotations of 7,755 different objects. The annotations are for every fifth frame in the videos. We train the model only with frames that have these annotations. We select a series of eight frames and apply random horizontal flipping of the frames and random reversing of the video to augment the data. We randomly select an object if there is more then one object in the first frame. We use Adam [3] optimizer with an initial learning rate of 0.00001. Our model converges in 120 epochs. Lambda, the regularization parameter, in the loss function is set to 0.05.

For inference, the offline-trained model is able to generate segmentation results for new videos without any online learning.

Table 1 displays the quantitative results on the YoutubeVOS validation set. J region similarity and F contour accuracy, described in [6] are shown. The unseen and seen scores are the scores for the object categories that are seen in the training set and not included in the data set respectively. The baseline model is trained on the full pixel-level annotations. The loss calculated at the pixel-level had the best results among the proposed methods.

Qualitative results are shown in table 2. In the first example with the tennis player, the proposed method tended to over segment the player. We hypothesize that the super-pixels generated were too large to identify just the player herself. This suggests that increasing the number of super-pixels will lead to better quality results so that each individual super-pixel is identifying a smaller region of the frame and will be able to identify smaller objects.
<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>J Seen</th>
<th>J Unseen</th>
<th>F Seen</th>
<th>F Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.31</td>
<td>0.42</td>
<td>0.25</td>
<td>0.34</td>
<td>0.21</td>
</tr>
<tr>
<td>Pixel-Level</td>
<td>0.15</td>
<td>0.19</td>
<td>0.14</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Super-Pixel-Level</td>
<td>0.14</td>
<td>0.19</td>
<td>0.11</td>
<td>0.16</td>
<td>0.09</td>
</tr>
<tr>
<td>Super-pixel w/Alpha</td>
<td>0.13</td>
<td>0.18</td>
<td>0.10</td>
<td>0.16</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 1. Baseline Model is trained with full Annotations. Pixel-level loss is the proposed loss calculated at the pixel level. Super-pixel Loss is the proposed loss calculated at the super-pixel-level.

Table 2. Comparison of qualitative results for the different models.

4. Conclusion

These results are preliminary and further research and training is needed to determine the efficacy of the proposed loss and method. It is possible that background super-pixels that are correctly labeled as background will have a high similarity with the foreground super-pixels and be mis-classified. Also in situations with multiple of the same object, the super-pixels of these objects will all have high similarity and can be mislabeled as one object. Further research needs to be done to handle these scenarios. In addition, using the full 30 frames per second instead of only frames with annotations will improve the overall results for all the models.

References


