Efficient Object Localization and Segmentation in Weakly Labeled Videos

Mrigank Rochan and Yang Wang

Department of Computer Science, University of Manitoba, Canada
{mrochan,ywang}@cs.umanitoba.ca

Abstract. In this paper, we tackle the problem of efficiently segmenting objects in weakly labeled videos. Internet videos (e.g., YouTube) are often associated with a semantic tag describing the main object within the video. However, this tag does not provide any spatial or temporal information about the object within the video. So these videos are weakly labeled. We propose a novel and efficient approach to localize the object of interest within the video and perform pixel-level segmentation. Given a video with an object tag, our proposed method automatically localizes the object and segments it from the background in each frame of the video. Our method combines object appearance modeling and temporal consistency among frames in a principled framework. Our method does not require user inputs or object detectors, so it can be potentially applied to videos of any object categories. We evaluate our method on a dataset consisting of more than 100 video shots of 10 different object categories. Our experimental results show that our method outperforms other baseline approaches.

1 Introduction

Today we have access to an enormous amount of video content through video sharing websites like YouTube. These videos are often associated with textual descriptions, such as tags. These tags are created by users to provide some information about the visual content (e.g., main object) present in the video. The object tag tells us whether an object is present in the video, but it does not provide any spatial or temporal information to localize the object within the video. Thus these videos are weakly labeled. In this paper, we tackle the problem of segmenting the object of interest in weakly labeled videos. This line of research will play an important role in many tasks related to video understanding. For example, it can enhance the browsing experience of users on video-sharing websites (e.g., YouTube). It can also improve video retrieval algorithms by removing the noisy videos or false positives from the search results.

Video segmentation is a fundamental problem in computer vision. Supervised learning of segmentation models requires all pixels in the training videos to be fully labeled. This is very time consuming and expensive. To address this drawback, weakly-supervised methods are proposed to alleviate the burden of labeling training video data. Weakly labeled videos have video-level labeling
instead of pixel-level labeling. For example, a video may have a video-level tag assigned to it, say “dog”. From this tag we can interpret that an object “dog” may be present in the video. However, we do not have any spatial or temporal information of the object “dog” within the video.

![Fig. 1. A demonstration of our proposed approach. Given an input video with an object tag, e.g. “dog” (1st row), our proposed method can localize (2nd row) and segment (3rd row) the object in each frame.](image)

Figure 1 demonstrates the pipeline of our proposed method. Given an input video with a tag, say “dog”, we want to localize the dog in each frame, and segment the pixels corresponding to “dog” from the background. Our method is generic and can be applied to videos of any object category.

## 2 Related Work

Video segmentation is an active area of research in computer vision. Some of the proposed approaches are unsupervised, e.g. region tracking [1], hierarchical graph model [23], multiple hypothesis tracking [4] and spatio-temporal based segmentation [5,6]. Unsupervised methods can only perform low-level video segmentation and can not provide semantic labels for the segments.

Supervised methods have also been studied for semantic video segmentation, e.g. [7]. The major drawback of supervised video segmentation is that it requires lots of labeled video data for training. To address this issue, semi-supervised video segmentation [8,9] methods are proposed. These methods address the limitation of supervised methods to some extent, but getting sufficient pixel-level labeled data is still nontrivial.

Weakly supervised video segmentation methods [10,11,12] are proposed to curtail the need of pixel-level labeled training video data. Our proposed method is inspired by this line of research. These methods use the semantic tags associated with the videos and do not require pixel-level labels. Since it is much easier to tag a video than labeling every pixels in the video, the labeling effort required is greatly reduced for weakly supervised methods. Rochan et al. [12] used video specific appearance model to localize object of interest in the video. Tang et al. [13] incorporated a temporal consistency model to their framework.