A Survey of Video Segmentation

Youcheng Li[1]

[1]**Xi’an Jiaotong University**

ABSTRACT

Video Segmentation (VS) aims at segmenting and tracking every pixel of input video clips. As a fundamental technique to scene understanding, it has received increasing attention in recent years due to its wide applications in many vision systems including autonomous driving and robot navigation. By definition, VS is an extension of Panoptic Segmentation (PS) into video domian with the goal of unifying Video Semantic Segmentation (VSS), and Video Instance Segmentation (VIS) into a single task. In this survey, we first introduce the classification of video segmentation tasks and focus on the introduction of low computation video segmentation task algorithms. After that, we enumerate the datasets and performance metrics for the video segmentation task.

INTRODUCTION

Video Segmentation (VS) aims at segmenting and tracking every pixel of input video clips. As a fundamental technique to scene understanding, it has received increasing attention in recent years due to its wide applications in many vision systems including autonomous driving and robot navigation. By definition, VS is an extension of Panoptic Segmentation (PS) into video domian with the goal of unifying Video Semantic Segmentation (VSS), and Video Instance Segmentation (VIS) into a single task.

Existing studies for video segmentation can be mainly divided into top-down methods and bottom-up approaches. Top-down methods try to solve VS as a multi-task learning problem via performing semantiv segmentation, instance segmentation and multiple object tracking individually. VPSNet is proposed to learn to fuse and warp features. In particular, it applies an optical flow network to align features and attention modules to fuse features. However, these approaches involve many hyper-parameters and ad-hoc designs, leading to a complex system. Bottom-up methods first perform semantic segmentation and then predict near frames’ centers to localize and track each instance where both features are connected by an ASPP module. Despite being simpler than top-down approaches, they rely on several post-processing components (i.e., NMS for center grouping, offline mask tracking).

Recently, inspired by the design of object queries in Detection Transformer (DETR), new efforts emerge to explore unified solutions to image panoptic segmentation. In particular, MaskFormer and K-Net unify ’things’ and ’stuff’ segmentation by dynamic kernels within the mask classification paradigm. The aforementioned studies in image segmentation motivate us to simplify cumbersome pipelines in video segmentation. It is noteworthy that several studies in Multi Object Tracking (MOT) also adopt the query design to achieve end-to-end learning. However, most of them introduce extra tracking queries or boxes to handle each tracked instance.

LOW COMPUTATION VIDEO SEGMENTATION ALGORITHM

However, those top-down methods are usually slow in speed, resulted from the multiple sequential processes in the pipeline. On the other hand, bottom-up methods naturally resolve the conflict by predicting non-overlapping segments. Only few works adopt the bottom-up approach, which typically starts with a semantic segmentation prediction followed by grouping operations to generate instance masks. Tackling panoptic segmentation in such a sequential order allows a simple and fast scheme, such as majority vote to merge semantic and instance segmentation results. Although obtaining promising fast inference speed, bottom-up approaches still demonstrate inferior performance compared to top-down ones prevailing in public benchmarks.

The difficulties faced by top-down methods, and the dearth of previous investigations into complementary approaches motivate us establish a simple, strong, and fast bottom-up baseline for panoptic segmentation.

DATASET AND PERFORMANCE METRIC

Video Segmentation Dataset

Video Object segmentation is a binary labeling problem aiming to separate foreground object(s) from the background region of a video. Click the following link to get detailed information about tasks and datasets:

- Task: Video Object Segmentation
- Task: Video Segmentation

Metric Analysis

A detailed analysis of existing metrics is essential for developing a new one. We first look into basic components PQ and MOTSA, before diving into the VPQ and PTQ metrics. We then condense these insights into properties that a good metric should satisfy.

Panoptic Quality (PQ)

Kirillove et al. proposed PQ to measure panoptic segmentation results. The set of true positive (TP_c), false positive (FP_c) and false negatives (FN_c) segments, for a particular segmantic class c, is formed by matching predictions p to ground-truth g based on the IoU scores. A threshold of greater than 0.5 IoU is chosen to guarantee unique matching. The overal metric is

$$PQ_c = \frac{\sum_{(p,g)\in TP_c} IoU(p,g)}{|TP_c| + \frac{1}{2}|FP_c| + \frac{1}{2}|FN_c|}$$

Defining a predicted segment with IoU of 0.51 as true positice and a segment with IoU of 0.49 as false positive (and the ground-truth segment as false negative) does not align well with human perception. Moreover, the metric is highly sensitive to false positives that cover few pixels, which is why most, if not all, panoptic segmentation methods use an additional post-processing step to remove stuff classes with few pixels or add 'void' class predictions to increase the PQ score.

A metric should be insenitive to such post-processing, and consider all predictions without requiring a threshold to define true positives on a segment level.

MOTA/MOTSA

Multi-Object Tracking Accuracy (MOTA) introduced the concept of ID switches $IDSW_c$ between frames of the prediction w.r.t. the ground-truth GT_c. It became the standard in many tracking benchmarks. Its derivative Multi-Object Tracking and Segmentation Accuracy (MOTSA) additionally considers segmentation, by matching objects based on segment overlap:

$$MOTSA_c = \frac{|TP_c| - |FP_c| - |IDSW_c|}{|GT_c|}$$

MOTA and MOTSA were analyzed exhaustively by HOTA: A Higher Order Metric for Evaluating Multi-object Tracking . The most relevant relevant drawbacks are:

- D1. The metric penalizes ID recovery, i.e., correcting mistakes gives worse scores.
- D2. The metric is unbounded and can take negative values, making scores hard to interpret.
- D3. The tracking precision is not considered, only recall is for the link above for details.
- D4. For detection, precision is much more important than recall, leading to a high imbalance.
- D5. Similar to PQ, MOTSA uses the threshold-based matching to define the set of true positives.

A metric should consider both precision and recall, and should not penalize correcting mistakes.

VPQ

Video Panoptic QUality (VPQ) for video panoptic segmentation is based on PQ and tailored to the sparsely annotated Cityscapes-VPS dataset. VPQ computes the average quality w.r.t. single frames and small spans of frames, by using a 2D and 3D IoU for matching, respectively:

$$VPQ = \frac{1}{K}\sum_k \frac{1}{N_{N_c lasses}} \sum_c \frac{\sum_{(p,g)\in TP_c^k IoU_{2D/3D}(p,g)}}{|TP_c^k| + \frac{1}{2}|FP_c^k| + \frac{1}{2}|FN_c^k|}$$

We note that these spans are constructed from every 5th video frame. Considering this averaging and that at most $K = 4$ frames are taken into account, the metric puts much more emphasis on segmentation than on association. When using more than 4 frames, the difficulty of the 3D IoU matching increases significantly. We agree with that assessment, but argue that we should by-pass threshold-based matching and propose a metric defined on the pixel level. When considering a setting like ours in which there are more stuff classes than countable things classes, VPQ reduces this task almost to video semantic segmentation as the metric is averaged over all classes. As a consequence, the importance of association varies with the dataset class split into stuff and things.

A metric should be suitable for evaluating not only short-term but also long-term tracks as well as give equal importance to segmentation and tracking, as required for real-world scenarios.

(s)PTQ

Hurtado et al. propose the Panoptic Tracking Quality (PTQ) measure:

$$PTQ_c = \frac{\sum_{(p,g)\in TP_c} IoU_{2D}(p,g) - |IDSW_c|}{|TP_c| + \frac{1}{2}|FP_c| + \frac{1}{2}|FN_c|}$$

The metric combines PQ and MOTSA, by changing the numerator of PQ to subtract the number of ID switches $IDSW_c$. The metric is computed in 2D on a frame by frame basis, thus is incapable of looking beyond first-order association mistakes. Also, PQ's original threshold-based matching is applied. Following this design, PTQ inherits the drawbacks from PQ as well as most issues from MOTSA as described above.

A metric should evaluate videos beyond single frames and decouple segmentation and association.

Metric Requirements

Given these insights, we argue for a change of paradigms. In detail, we define the following properties that a metric should satisfy:

- P1. Analyze full videos at pixel level: The metric should work on full videos and at the pixel level rather than on single frames or at segment level.

- P2. No threshold-based matching: The metric should treat all pixels equally in space and time.

- P3. No penalty for mistake correction: Correcting errors in ongoing tracks should not be penalized, but encouraged to obtain long-term track consistency.

- P4. Consider precision and recall for association: For association errors (e.g., ID transfer) the metric should take precision and recall into account.

- P5. Decouple errors: The metric could be decoupled into components, enabling detailed analysis for different aspects of performance.

STQ Metric

The previous discussions motivate us to propose the Segmentation and Tracking Quality (STQ).

Formal Defination

Our benchmark requires a mapping f of every pixcel (x,y,t) (indexed by spatial coordinates (x,y) and frame t) of a video Ω to a semantic class c and a track ID id. We denote the ground-truth as $gt(x,y,t)$ and the prediction as $pr(x,y,t)$. STQ combines Association Quality (AQ) and Segmentation Quality (SQ) that measure the tacking and segmentation quality, respectively.

Association Quality (AQ)

The proposed AQ is designed to work at a pixel-level of a full video (c.f., P1). Importantly, all correct and incorrect associations influence the score, independent of whether a segment is above or below an IoU threshold (c.f., P2). We define the prediction for a particular id as:

$$pr_{id}(id) = \{(x,y,t) | pr(x,y,t) = (c,id), c \in C^{th}\}$$

here we only consider ”trackable” objects C^{th}, i.e., things. The ground-truth is defined analogously. Mislabeling the semantic class of a pixel is not penalized in AQ, and will only be judged by segmentation quality described later. This decoupling prevents penalizing wrong semantic predictions twice. We also do not require a consistent semantic class label per track, for reasons illustrated in the following example: A ground-truth van track can easily be mistaken as a car, which should be penalized in the segmentation score, but not in the association score. Moreover, once it becomes overall increased score for our proposed metric. However, when requiring on semantic class per track, this prediction would be split into one track for car and one for van. As a result, this split that corrects the semantic class would receive a lower score than a prediction that keeps the wrong class, which contradicts P3. We call this effect class-recovery in line with the tracking term ID recovery.

We define the true positive associations (TPA) of an specific ID as:

$$TPA(p,g) = |pr_{id}(p) \cap gt_{id}(g)|$$

Similarly, false negative associations (FNA) and false positive associations (FPA) can be defined to compute precision P_{id} and recall R_{id} (c.f., P4). Increasing precision requires the minimization of FPAs, while increasing recall requires minimization of FNAs. We define the IoU_{id} for AQ as follows:

$$IoU_{id}(p,g) = \frac{P_{id}(p,g) \times R_{id}(p,g)}{P_{id}(p,g) + R_{id}(p,g) - P_{id}(p,g) \times R_{id}(p,g)}$$

Following our goal of lone-term track consistency, we encourage ID recovery by weighting the score of each predicted tube by their TPA. Without this weighting, a recovered ID would not achieve a higher score. In total, the association quality (AQ) is defined as follows.

$$AQ = \frac{1}{|gt\_tracks|} \sum_{g \in gt\_tracks} \frac{1}{|gt_{id}(g)|} \sum_{p, |p \cap q| \neq \emptyset} TPA(p,g) \times IoU_{id}(p,g)$$

For each ground-truth track g, its association score $AQ(g)$ is normalized by its size rather than by the sum of all TPA, which penalizes the removal of correct segments with wrong IDs noticeably.

Segmentation Quality (SQ)

In semantic segmentation, Intersection-over-Union (IoU) is the most widely adopted metric. As IoU only considers the semantic labels, it fulfills the property of decoupling segmentation and association (c.f., P5). Additionally, it allows us to measure the quality for crowd regions. Formally, given $pr(x,y,t)$, and class c we define:

$$pr_{sem}(c) = \{(x,y,t) | pr(x,y,t) = (c,*)\}$$

The ground-truth is defined analogously. We then define SQ to be the mean IoU score:

$$SQ = mIoU = \frac{1}{|C|} \sum_{c \in C} \frac{|pr_{sem}(c) \cap gt_{sem}(c)|}{|pr_{sem}(c) \cup gt_{sem}(c)|}$$

Segmentation and Tracking Quality (STQ)

The overall STQ score is the geometric mean:

$$STQ = (AQ \times SQ)^{\frac{1}{2}}$$

The geometric mean is preferable over the arithmetic mean, as it rewards joint segmentation and tracking. Methods specializing in only one aspect of the task will therefore receive a lower score in this setting.

ACKNOWLEDGMENTS

Thanks to Professor Liwei Wang and Dr. Dong Wang for their support.

PAPER LIST

Click This Link to Get Paper List of Video Segmentation