Computer Vision and Image Understanding 143 (2016) 159–172

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Primary object discovery and segmentation in videos via graph-based transductive inference

Huiling Wang [a], Tinghuai Wang [b,*]

[a] Lappeenranta University of Technology, Lappeenranta 53850, Finland
[b] Nokia Technologies, Visiokatu 3, Tampere 33720, Finland

ARTICLE INFO

Article history:
Received 2 October 2014
Accepted 15 November 2015

Keywords:
Graph-based transductive inference
Video object segmentation
Object proposal

ABSTRACT

The proliferation of video data makes it imperative to develop automatic approaches that semantically analyze and summarize the ever-growing massive visual data. As opposed to existing approaches built on still images, we propose an algorithm that detects recurring primary object and learns cohort object proposals over space-time in video. Our core contribution is a graph transduction process that exploits both appearance cues learned from rudimentary detections of object-like regions, and the intrinsic structures within video data. By exploiting the fact that rudimentary detections of recurring objects in video, despite appearance variation and sporadity of detection, collectively describe the primary object, we are able to learn a holistic model given a small set of object-like regions. This prior knowledge of the recurring primary object can be propagated to the rest of the video to generate a diverse set of object proposals in all frames, incorporating both spatial and temporal cues. This set of rich descriptions underpins a robust object segmentation method against the changes in appearance, shape and occlusion in natural videos. We present extensive experiments on challenging datasets that demonstrate the superior performance of our approach compared with the state-of-the-art methods.

1. Introduction

Segmenting object from video remains an open challenge with recent advances relying upon prior knowledge supplied via interactive initialization or correction [1–6]. Yet fully automatic video object discovery and segmentation [7–12] remains useful in scenarios where the human in the loop is impractical, such as video summarization or ingest pre-processing for video indexing or recognition. This is a very challenging task due to the lack of prior knowledge about object appearance, shape or position. Furthermore, variance in illumination and occlusion relationships introduce ambiguities that in turn induce instability in boundaries and the potential for localized under- or over-segmentation.

This paper proposes a novel automatic primary video object discovery and segmentation algorithm in which the segmentation of each frame is driven by set of rich object models learned from *spatio-temporally dense and coherent object proposals*. Following [9–14], the primary video object refers to the object that presents saliently, in terms of either appearance or motion, in most of the frames. The core novel contribution is our *graph transduction* approach to the efficient learning of the dense video object proposals which enables the detection and segmentation of objects in complex dynamic scenes without suffering from appearance variation or object occlusion over time. In contrast to previous techniques, our algorithm learns and extracts object proposals from scratch to account for the evolution of object's appearance, shape and location with time, as opposed to selecting from existing per-frame detections of object-like regions [9–13].

Our strategy is to create feature-based rudimentary detections of regions for the primary object by learning from weakly labelled examples of object-like regions. These detections serve as informative indicators of the appearance and location of the object. We propagate this learned prior knowledge on an undirected space-time graph consisting of regions, solving the transduction learning efficiently with a fast convergence technique [15]. Inference at the region level further makes our dense video object proposal extraction approach a practical solution for automatic object segmentation on natural video sequences.

We describe our proposed video object proposal algorithm in Section 3, presenting the utilization of video object proposals for robust video object segmentation in Section 4. In Section 6, we evaluate our video object proposal and segmentation approach on benchmark dataset and additional dataset comprising challenging video clips exhibiting clutter, occlusion and agile motion, comparing against state-of-the-art semi-automatic and automatic algorithms.

* Corresponding author.
E-mail addresses: huiling.wang@lut.fi (H. Wang), tinghuai.wang@gmail.com, tinghuai.wang@nokia.com (T. Wang).

http://dx.doi.org/10.1016/j.cviu.2015.11.006
1077-3142/

2. Related work

Generic object detection has been intensively studied in context of still images recently [16–21]. Alexe et al. [16] introduced the objectness measure which computes the probability that a window contains any object, using a Bayesian classifier based on multiple cues. Carreira et al. [17] (CPMC) proposed to use several graphcuts running using random positive and negative seeds. Each generated foreground mask serves as an object proposal, and the proposals are ranked according to a learned scoring function. Similarly to CPMC, Endres and Hoiem [18] proposed to generate multiple foreground segmentations and use these as object proposals using binary CRF segmentation with random seeds. [19] performs an ad-hoc hierarchical bottom-up agglomeration of groups of regions and a fixed number of proposals are generated at each step of the agglomeration. Manen et al. [20] proposed an approach based on randomly growing groups of regions, which allows to generate any desirable number of object proposals. Cheng et al. [21] introduced a simple and fast objectness measure to compute the objectness of each image window at various scale and aspect ratio. The bounding box based objectness measure methods [16,21] share the similar limitations that a bounding box might not localize the object instances as accurately as a segmentation region. Generating object proposals incorporating temporal information has been receiving more attentions recently [22,23]. Sharir and Tuytelaars [22] proposed to extract object proposals in each frame separately which are linked across frames into object hypotheses. This approach suffers from the mis-segmentations of object proposals in each independent frame. Oneata et al. [23] proposed a supervoxel method for spatio-temporal detection. However, supervoxel based approaches usually become computationally infeasible for pixel counts in even moderate size videos, and often undersegment small or fast moving objects that form disconnected space-time volumes.

Our method follows the segmentation based approach to generating video object proposals which provides a set of rich descriptions underpinning robust segmentation and many other applications against large variations of appearance, shape and occlusion in natural videos. As apposed to those image based generic object detection algorithms which typically generate an excessive amount of proposals ($> 10^4$), our approach generates cohort object proposals over space-time to capture the essential parts of tentative primary object exploring cues beyond the single still image.

Video object segmentation methods requiring user to provide an initial annotation of the first frame have been proposed, which either propagate the annotation to drive the segmentation in successive frames [1–6,24] or perform spatio-temporal grouping [25,26]. The former group of methods heavily rely on motion estimates and may fail in segmenting videos with complex motions or varying object appearance. Although stability is achieved in the latter methods, they usually become computationally infeasible for pixel counts in even moderate size videos, and often fail in dealing with fast moving objects.

Automatic video object segmentation methods have also been proposed as a consequence of the prohibitive cost of user intervention in processing large amounts of video data in most computer vision applications. Methods like [12,27–31] take a bottom-up approach based on spatio-temporal appearance and motion constraints. Motion segmentation methods [32–38] cluster pixels or regions in video employing long-term motion trajectories analysis, which require the motion of the primary object to be neither too similar with the background nor too fast. Occlusion boundary approach [37] has been proposed to detecting occlusion boundaries and assigning figure/ground labels to both sides of those boundaries. Layered models have been studied in [32,35,36,39,40]. Methods which generate over-segmentations for later processing analog to still-image regions [41] have also been proposed [42,43], by applying spatio-temporal clustering based on low level features. Papazoglou and Ferrari [12] determine an initial set of foreground pixels based purely on motion and refine the FG/BG labels using Graph Cut. However, without any top-down explicit notion of object, all of these automatic methods produce segmentations without corresponding to any particular object with semantic meaning.

Several recent methods [9–11,13,14] are proposed based on exploring recurring object-like regions from still images by measuring generic object appearance [18]. Lee et al. [9] proposed to extract 'key-segments' of the primary object by performing clustering in a pool of object proposals from each frame of the video. The weakness of this approach is that the object proposal pool combines regions across all frames and discards the spatial and temporal information of each region. Ma and Latecki [10] proposed to leverage the temporal information by utilizing binary appearance relation between regions in different frames and model the object region selection as a constrained Maximum Weight Cliques problem. Zhang et al. [11] improved this approach by introducing optical flow to track the evolution of object shape and appearance and solving the primary object proposal selection problem as the longest path problem for Directed Acyclic Graph (DAG). There are mainly two limitations with these later two approaches [10,11]. First, both approaches propose to select or merge per-frame extracted object-like regions based on the objectness score which is computed locally in each frame, regardless of the prior knowledge of the corresponding object learned from other frames; their performance heavily relies on the quality of the initial rudimentary detection of object-like regions which is highly unreliable in practice. The initial object proposals generated using [18] normally contain a large amount of erroneous regions. Second, both approaches assume all object-like regions within each frame are independent and do not explicitly consider spatial affinity. This substantially limits the size of the object proposal especially when the primary object is comprised of multiple regions with distinct appearances. An additional limitation of [11] is that it employs optical flow warped region overlap to merge object-like regions into a new region which may introduce further spurious proposals due to inherent motion estimate error. Li et al. [13] proposed to track a pool of figure-ground segments in each frame and incrementally to learn a long-term object appearance model. However the incrementally built appearance model heavily relies on greedy matching and also suffers from the cumulative motion estimation error. Yang et al. [14] proposed a method to fuse appearance and motion saliency maps for discovering primary video object. All the above methods do not build an explicit holistic appearance model but relies on local heuristics and motion for selecting and merging the object proposals or saliency maps.

To address the limitations of the above approaches [9–11,13,14], we propose to learn a holistic appearance model from the rudimentary detection of object-like regions across the whole video to drive the generation of dense object proposals. We propagate the prior knowledge from rudimentary detections on an undirected space-time graph consisting of regions by performing transduction learning, with respect to both low level cues collectively revealed by the appearance model and the intrinsic structure within video data. The transduction learning is guided by the initially detected evidence by collectively learning the initial sparse object-like regions, rather than directly using the local static 'objectness' score. Spatio-temporally coherent and dense object proposals are generated to facilitate robust object segmentation in challenging natural videos.

Our segmentation is driven by Markov Random Field (MRF) approach. A variety of MRF models as well as inference and learning methods have been developed for addressing numerous computer vision problems during the past decade. MRF models can be categorized into pairwise models and higher-order models. Various works have investigated the modeling of vision problems using pairwise MRFs (e.g., [36,44–48]) and the efficient inference in pairwise MRFs