ARTICLE IN PRESS
JID: YCVIU [m5G;June 24, 2016;12:49]

Computer Vision and Image Understanding 000 (2016) 1–9

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Segmentation models diversity for object proposals

Marco Manfredi[a,*], Costantino Grana[a], Rita Cucchiara[a], Arnold W.M. Smeulders[b]

[a] University of Modena and Reggio Emilia, Italy
[b] ISLA, Informatics Institute, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 27 November 2015
Revised 1 April 2016
Accepted 19 June 2016
Available online xxx

Keywords:
Segmentation
Supervised learning
Object proposals

ABSTRACT

In this paper we present a segmentation proposal method which employs a box-hypotheses generation step followed by a lightweight segmentation strategy. Inspired by interactive segmentation, for each automatically placed bounding-box we compute a precise segmentation mask. We introduce diversity in segmentation strategies enhancing a generic model performance exploiting class-independent regional appearance features. Foreground probability scores are learned from groups of objects with peculiar characteristics to specialize segmentation models. We demonstrate results comparable to the state-of-the-art on PASCAL VOC 2012 and a further improvement by merging our proposals with those of a recent solution. The ability to generalize to unseen object categories is demonstrated on Microsoft COCO 2014.

1. Introduction

Automatic object segmentation is among the oldest topics in computer vision, and apparently one of the hardest, in view of the results obtained thus far. Other topics, such as image recognition and image search, have increased from a poor to a solid performance in just a decade. While first ignoring location information altogether (Jégou et al., 2010; Perronnin et al., 2010; Uijlings et al., 2010), recognition and search have recently reintroduced locality where it now plays an important role (Carreira et al., 2012a; Uijlings et al., 2013). We can obtain object localization in the form of a set of box-hypotheses (Cheng et al., 2014; Zitnick and Dollár, 2014) or precise segmentation masks (Carreira and Sminchisescu, 2012; Endres and Hoiem, 2014; Krähenbühl and Koltun, 2014).

Inspired by interactive segmentation, where every object is perfectly inscribed in a user-placed bounding-box and then segmented, our goal is to start from a set of automatically obtained bounding-boxes and for each of them extract a precise segmentation (Weiss and Taskar, 2013). A clear problem with respect to the interactive segmentation setting is that the number of object candidates to analyze is in the order of 1000 per image and not only 1 per object, leading to large running times (Weiss and Taskar, 2013 reports 6 to 10 min per image). We aim to develop a method to refine box-hypotheses scalable to thousands of proposals.

As objects may be discriminated from the background on the basis of their edge information, their texture, or other appearance cues, it is unlikely that there exists one single model for generic object segmentation (Kuang et al., 2012; Malisiewicz and Efros, 2007). Differentiation and combination of several segmentation strategies is necessary to control object diversity (Uijlings et al., 2013). One extreme approach for diversity is to build a new segmentation model for each new class of objects (Dai and Hoiem, 2012; Weiss and Taskar, 2013). A recognition step is thus required to select the appropriate model. Class-specific segmentation models are hard to apply in large-scale applications (Lin et al., 2014), and they are by definition not applicable to an unknown class of objects. We use the progress in the field of segmentation to strive for a class-independent approach (Endres and Hoiem, 2014; Krähenbühl and Koltun, 2014), while introducing diversity in the segmentation strategy to enhance its generic performance where needed (Fig. 1).

Our approach starts with box-hypotheses built from edge statistics (Zitnick and Dollár, 2014). On the basis of lightweight superpixel features, we assess the probability of belonging to the foreground. The use of spatially-smooth visual features (e.g. geodesic distance) allows for accurate segmentations while avoiding any time-consuming regularization (Boykov and Jolly, 2001). Rather, we rely on a simple threshold of the foreground probabilities to generate the binary segmentations. We also avoid any proposal reranking (Carreira et al., 2012a; Endres and Hoiem, 2014) delegating the ranking to the stage of the box-hypotheses. These choices allow for a fast segmentation proposal generation.

During training, diversity is included by unsupervised clustering, sorting objects into different types on the basis of regional appearance features. Ideally, each cluster contains a specific group of objects suited for a specific segmentation approach. For each group

* Corresponding author.
E-mail address: marco.manfredi@unimore.it (M. Manfredi).

http://dx.doi.org/10.1016/j.cviu.2016.06.005
1077-3142/

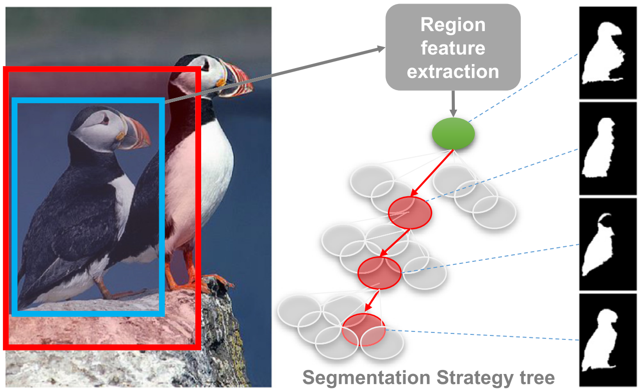

Fig. 1. Segmentation strategy diversification is employed to produce diverse proposals.

of objects, specialized segmentation models are learned. The same features are used to assign an unknown object to one of these clusters when applying the algorithm.

Our contributions are:

1. We propose a fast and class-independent segmentation technique, starting from recent methods for generating box-hypotheses;
2. By grouping objects into clusters, each suited for a specific segmentation strategy, we effectively achieve object-group diversity, reaching state-of-the-art results on PASCAL VOC 2012. We demonstrate how the learned segmentation strategies generalize to unseen categories on the Microsoft COCO 2014 dataset.
3. We further demonstrate a considerable improvement in segmentation accuracy over the state-of-the-art by enhancing the diversity after merging with a recent segmentation strategy (Krähenbühl and Koltun, 2014).

The objects clusters obtained while diversifying the segmentation models are also used to highlight when our method or (Krähenbühl and Koltun, 2014) are providing the best candidates. The highlight illustrates the importance of segmentation model diversity in the success of the integrated solution.

2. Related work

Object localization with candidate segmentations has attracted a lot of attention in the last years (Arbelaez et al., 2014; Carreira et al., 2012a; Endres and Hoiem, 2014; Kim and Grauman, 2012; Krähenbühl and Koltun, 2014; Malisiewicz and Efros, 2007; Ren and Shakhnarovich, 2013; Weiss and Taskar, 2013), mainly due to the improvement that precise localization offers in object recognition settings (Carreira et al., 2012b; Carreira and Sminchisescu, 2012).

The CPMC approach (Carreira et al., 2012a) uses multiple graph-cut computations at pixel-level to compute segmentation candidates from seeds placed on a grid over the image. The region level affinities proposed in Endres and Hoiem (2014) have inspired our foreground probability score. Differently from our work, however, in the reference they are computed on bigger regions and transferred to a superpixel graph regularized in a CRF.

The RIGOR approach (Humayun et al., 2014) tackles the problem of computing hundreds of graph-cut computations (as done in Carreira et al. (2012a)) modifying the graph structure. Their method is able to compute simultaneous segmentations for different seeds/potential, leading to a considerable speed up.

The approach in Kim and Grauman (2012) is based on the idea that objects of different categories have similar local shapes. As a consequence, masks can be transferred from other objects and slightly adapted to the object of interest.

The Geodesic Object Proposals technique (Krähenbühl and Koltun, 2014) is based on geodesic distances from automatically placed foreground and background seeds. Their method is able to compute spatially smooth segmentations without employing regularization techniques such as graph-cut, used for example in Kim and Grauman (2012). Avoiding the regularization step speeds up the segmentation considerably, and thus we adopt the same geodesic features in our method.

In Malisiewicz and Efros (2007) the importance of segmentation in object recognition is stressed, along with a numerical demonstration of the importance of differentiating among segmentation techniques. The technique presented in Arbelaez et al. (2014),Pont-Tuset et al. combines edge detection, hierarchical segmentation and object proposals based on region grouping. Selective Search (Uijlings et al., 2013) uses segmentation strategy diversification by changing the criterion on which adjacent regions are being merged. The diversification enlarges the search space for possible objects. Both (Weiss and Taskar, 2013) and (Ren and Shakhnarovich, 2013) use size as a cue to differentiate segmentation models, based on the idea that the relevance of visual features is related to object size. While (Weiss and Taskar, 2013) uses class-specific shape priors, (Ren and Shakhnarovich, 2013) only relies on class-independent probabilistic models. In order to diversify segmentation strategies without including class information, we leverage regional level features, including size, in a hierarchically structured decision model.

In Xia et al. (IEEE), bounding-boxes coming from an object detector along with segment hypothesis coming from CPMC are used to initialize a semantic segmentation algorithm. A shape guidance term is computed for each box and regularized with a graph-cut.

In the interactive segmentation approach presented in Kuang et al. (2012), segmentation models are adapted to each object using two manually traced polygons to learn the optimal parameters of the segmentation model (e.g. feature importance). Our solution strives to a similar specialization in an automatic setting.

3. From bounding boxes to segmentation masks

Starting from a bounding-box R we want to outline the contained object. Locality in segmentation is of fundamental importance, and thus only a close neighborhood of the object is considered in the segmentation process.

We assume that the object is fully contained in R, by labeling the outside region as background. The area surrounding R, obtained by enlarging its area by a 50% factor, defines the background area (used to model background information). We further assume that the center of R belongs to the object, using it as the foreground seed (Fig. 2).

A superpixel over-segmentation of the image is computed, and each superpixel is labeled according to the area of maximum overlap. For each box proposal we obtain two sets of superpixels: the background seeds $\mathcal{B}$ and the foreground seed $\mathcal{F}$ (i.e. the superpixel containing the center of R).

A set of features (9 in total), presented below, is extracted from each superpixel and used in a supervised setting to compute a foreground probability score.

From $\mathcal{F}$ and $\mathcal{B}$ two color histograms are extracted representing the RGB color distributions of foreground and background (C_f and C_b respectively). For each superpixel S_i, we compute the similarity of its color histogram C_{S_i} with respect to C_f and C_b, and the difference between the two.

The geodesic distance to foreground and background seeds is another important feature of our framework. Following (Krähenbühl and Koltun, 2014), a graph over the superpixel