DOI 10.1007/s11263-012-0574-z

Superparsing

Scalable Nonparametric Image Parsing with Superpixels

Joseph Tighe · Svetlana Lazebnik

Received: 10 May 2012 / Accepted: 11 September 2012 / Published online: 6 October 2012

Abstract This paper presents a simple and effective nonparametric approach to the problem of image parsing, or labeling image regions (in our case, superpixels produced by bottom-up segmentation) with their categories. This approach is based on lazy learning, and it can easily scale to datasets with tens of thousands of images and hundreds of labels. Given a test image, it first performs global scene-level matching against the training set, followed by superpixel-level matching and efficient Markov random field (MRF) optimization for incorporating neighborhood context. Our MRF setup can also compute a simultaneous labeling of image regions into semantic classes (e.g., tree, building, car) and geometric classes (sky, vertical, ground). Our system outperforms the state-of-the-art nonparametric method based on SIFT Flow on a dataset of 2,688 images and 33 labels. In addition, we report per-pixel rates on a larger dataset of 45,676 images and 232 labels. To our knowledge, this is the first complete evaluation of image parsing on a dataset of this size, and it establishes a new benchmark for the problem. Finally, we present an extension of our method to video sequences and report results on a video dataset with frames densely labeled at 1 Hz.

Keywords Scene understanding · Image parsing · Image segmentation

1 Introduction

This paper addresses the problem of image parsing, or segmenting all the objects in an image and identifying their categories. Many approaches to this problem have been proposed recently, including ones that estimate labels pixel by pixel (He et al. 2004; Ladicky et al. 2010; Shotton et al. 2006, 2008), ones that aggregate features over segmentation regions (Galleguillos et al. 2010; Gould et al. 2009; Hoiem et al. 2007; Malisiewicz and Efros 2008; Rabinovich et al. 2007; Socher et al. 2011), and ones that predict object bounding boxes (Divvala et al. 2009; Felzenszwalb et al. 2008; Heitz and Koller 2008; Russell et al. 2007). Most of these methods operate with a few pre-defined classes and require a generative or discriminative model to be trained in advance for each class (and sometimes even for each training exemplar (Malisiewicz and Efros 2008, 2011)). Training can take days and must be repeated from scratch if new training examples or new classes are added to the dataset. In most cases (with the notable exception of Shotton et al. 2008), processing a test image is also quite slow, as it involves steps like running multiple object detectors over the image, performing graphical model inference, or searching over multiple segmentations.

While most existing methods are tailored for "closed universe" datasets, a new generation of "open universe" datasets is beginning to take over. An example open-universe dataset is LabelMe (Russell et al. 2008), which consists of complex, real-world scene images that have been segmented and labeled by multiple users (sometimes incompletely or noisily). There is no pre-defined set of class labels; the dataset is constantly expanding as people upload new photos or add annotations to current ones. In order to cope with such datasets, vision algorithms must have much faster training and testing times, and they must make it easy to

J. Tighe (✉) · S. Lazebnik
Computer Science Department, University of North Carolina,
Chapel Hill, NC, USA
e-mail: jtighe@cs.unc.edu

S. Lazebnik
e-mail: lazebnik@cs.unc.edu