Pattern Recognition Letters 72 (2016) 4–14

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

The three R's of computer vision: Recognition, reconstruction and reorganization☆

Jitendra Malik[a,*], Pablo Arbeláez[b], João Carreira[a], Katerina Fragkiadaki[a], Ross Girshick[c], Georgia Gkioxari[a], Saurabh Gupta[a], Bharath Hariharan[c], Abhishek Kar[a], Shubham Tulsiani[a]

[a] EECS, UC Berkeley, Berkeley, CA 94720, USA
[b] Universidad de los Andes, Bogotá 111711, Colombia
[c] Facebook, Seattle, WA 98101, USA

ARTICLE INFO

Article history:
Available online 8 February 2016

Keywords:
Object recognition
Action recognition: grouping
Segmentation
3D models
Shape reconstruction

ABSTRACT

We argue for the importance of the interaction between recognition, reconstruction and re-organization, and propose that as a unifying framework for computer vision. In this view, recognition of objects is reciprocally linked to re-organization, with bottom-up grouping processes generating candidates, which can be classified using top down knowledge, following which the segmentations can be refined again. Recognition of 3D objects could benefit from a reconstruction of 3D structure, and 3D reconstruction can benefit from object category-specific priors. We also show that reconstruction of 3D structure from video data goes hand in hand with the reorganization of the scene. We demonstrate pipelined versions of two systems, one for RGB-D images, and another for RGB images, which produce rich 3D scene interpretations in this framework.

1. Introduction

The central problems in computer vision are recognition, reconstruction and reorganization (Fig. 1).

Recognition is about attaching semantic category labels to objects and scenes as well as to events and activities. Part–whole hierarchies (partonomies) as well as category–subcategory hierarchies (taxonomies) are aspects of recognition. Fine-grained category recognition includes as an extreme case instance level identification (e.g. Barack Obama's face).

Reconstruction is traditionally about recovering three-dimensional geometry of the world from one or more of its images. We interpret the term more broadly as "inverse graphics" – estimating shape, spatial layout, reflectance and illumination – which could be used together to render the scene to produce an image.

Reorganization is our term for what is usually called "perceptual organization" in human vision; the "re" prefix makes the analogy with recognition and reconstruction more salient. In computer vision the terms grouping and segmentation are used with approximately the same general meaning.

Mathematical modeling of the fundamental problems of vision can be traced back to geometers such as Euclid [10], scientists such as Helmholtz [41], and photogrammetrists such as Kruppa [51]. In the twentieth century, the Gestaltists led by Wertheimer [85] emphasized the importance of perceptual organization. Gibson [26] pointed out the many cues which enable a moving observer to perceive the three-dimensional structure of the visual world.

The advent of computers in the middle of the twentieth century meant that one could now develop algorithms for various vision tasks and test them on images, thus creating the field of computer vision. [64] is often cited as the first paper in this field, though there was work on image processing and pattern recognition even before that. In recent years, progress has been very rapid, aided not only by fast computers, but also large annotated image collection such as ImageNet [16].

But is there a unifying framework for the field of computer vision? If one looks at the proceedings of a recent computer vision conference, one would notice a variety of applications using a wide range of techniques such as convex optimization, geometry, probabilistic graphical models, neural networks, and image processing. In the early days of computational vision, in the 1970s and 1980s, there was a broad agreement that vision could be usefully broken up into the stages of low level, mid level and high level vision. [58] is perhaps the best articulation of this point of view, with low level vision corresponding to processes such as edge detection, mid

☆ This paper has been recommended for acceptance by Rama Chellapp.

* Corresponding author. Tel.: +1 510 642 7597.
E-mail address: malik@cs.berkeley.edu, malik@eecs.berkeley.edu (J. Malik).

http://dx.doi.org/10.1016/j.patrec.2016.01.019
0167-8655/

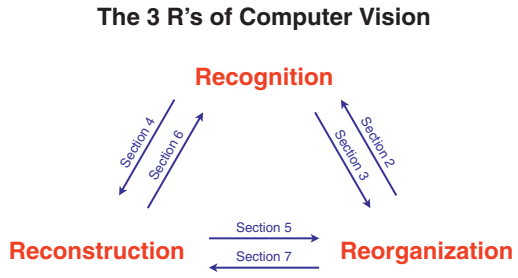

Fig. 1. The 3R's of vision: recognition, reconstruction and reorganization. Each of the six directed arcs in this figure is a useful direction of information flow.

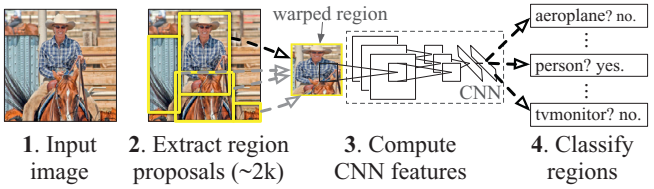

Fig. 2. R-CNN – Region-based Convolutional Network: object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional network (CNN), and then (4) classifies each region using class- specific linear SVMs. We trained an R-CNN that achieves a mean average precision (mAP) of 62.9% on PASCAL VOC 2010. For comparison, [78] report 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part models perform at 33.4%. On the 200-class ILSVRC2013 detection dataset, we trained an R-CNN with a mAP of 31.4%, a large improvement over OverFeat [67], which had the previous best result at 24.3% mAP.

level vision leading to representation of surfaces, and high level vision corresponding to object recognition. The process was thought of a being primarily feed-forward and bottom up. In the 1990s, this consensus gradually dissipated. Shape-from-X modules, with the exception of those based on multiple view geometry, proved to be not robust for general images, so the bottom up construction of Marr's desired 2.5D sketch proved infeasible. On the other hand, machine learning approaches to object recognition based on sliding windows started to succeed on real world images e.g. Viola and Jones' [80] work on face detection, and these didn't quite fit Marr's paradigm.

Back in the 1990s, one of us, [57] argued that grouping and recognition ought to be considered together. Bottom up grouping could produce candidates for consideration by a recognition module. Another paper from our group, [63] advocated the use of superpixels for a variety of tasks. This got some traction e.g. multiple segmentation hypotheses were used by Hoiem et al. [42] to estimate the rough geometric scene structure and by Russell et al. [66] to automatically discover object classes in a set of images, and Gu et al. [32] showed what were then state of the art results on the ETH-Z dataset. But the dominant paradigm remained that of sliding windows, and the state of the art algorithms on the PASCAL VOC challenge through 2011 were in that paradigm.

This has changed. The "selective search" algorithm of [78] popularized the multiple segmentation approach for object detection by showing strong results on PASCAL object detection. EdgeBoxes [88] outputs high-quality rectangular (box) proposals quickly (~0.3 s per image). Other methods focus on pixel-wise segmentation, producing regions instead of boxes. Top performing approaches include CPMC [12], RIGOR [45], MCG [4], and GOP [49]. For a more in-depth survey of proposal algorithms, [43] provide an insightful meta-evaluation of recent methods.

In this paper we propose to go much further than the link between recognition and reorganization. That could be done with purely 2D reasoning, but surely our final percept must incorporate the 3D nature of the world? We will highlight a point of view that one of us (Malik) has been advocating for several years now, that instead of the classical separation of vision into low level, mid level and high level vision, it is more fruitful to think of vision as resulting from the interaction of three processes: recognition, reconstruction and reorganization which operate in tandem, and where each provides input to the others and fruitfully exploits their output. We aim for a grand unified theory of these processes, but in the immediate future it may be best to model various pairwise interactions, giving us insight into the representations that prove most productive and useful. In the next six sections, we present case studies which make this point, and we conclude with a pipeline which puts the different stages together.

Note that the emphasis of this paper is on the relationship between the 3R's of vision, which is somewhat independent of the (very important) choice of features needed to implement particular algorithms. During the 1970s and 1980s, the practice of computer vision was dominated by features such as edges and corners which offered the benefit of massive data compression, a necessity in a time when computing power was orders of magnitude less than today. The community moved on to the use of linear filters such as Gaussian derivatives, Gabor and Haar wavelets in the 1990s. The next big change was the widespread use of histogram based features such as SIFT [55] and HOG [15]. While these dominated for more than a decade, we are now completing yet another transition, that to "emergent" features from the top layers of a multilayer convolutional neural network [53] trained in a supervised fashion on a large image classification task. Neural networks have proved very compatible to the synthesis of recognition, reconstruction and reorganization.

2. Reorganization helps recognition

As noted earlier, the dominant approach to object detection has been based on sliding-window detectors. This approach goes back (at least) to early face detectors [79], and continued with HOG-based pedestrian detection [15], and part-based generic object detection [20]. Straightforward application requires all objects to share a common aspect ratio. The aspect ratio problem can be addressed with mixture models (e.g. [20]), where each component specializes in a narrow band of aspect ratios, or with bounding-box regression (e.g. [20,67]).

The alternative is to first compute a pool of (likely overlapping) image regions, each one serving as a candidate object, and then to filter these candidates in a way that aims to retain only the true objects. By combining this idea with the use of convolutional network features, pretrained on an auxiliary task of classifying ImageNet, we get the Region-based Convolutional Network (R-CNN) which we describe next.

At test time, R-CNN generates around 2000 category-independent region proposals for the input image, extracts a fixed-length feature vector from each proposal using a convolutional neural network (CNN) [53], and then classifies each region with category-specific linear SVMs. We use a simple warping technique (anisotropic image scaling) to compute a fixed-size CNN input from each region proposal, regardless of the region's shape. Fig. 2 shows an overview of a Region-based Convolutional Network (R-CNN) and Table 1 presents some of our results.

R-CNNs scale very well with the number of object classes to detect because nearly all computation is shared between all object categories. The only class-specific computations are a reasonably small matrix-vector product and greedy non-maximum suppression. Although these computations scale linearly with the number of categories, the scale factor is small. Measured empirically, it takes only 30 ms longer to detect 200 classes than 20 classes on a CPU, without any approximations. This makes it feasible to