Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Vision and Image Understanding 100 (2005) 41–63
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Selective visual attention enables learning and recognition of multiple objects in cluttered scenes

Dirk Walther [a,*,1], Ueli Rutishauser [a,1], Christof Koch [a,b], Pietro Perona [a,c]

[a] *Comput. and Neural Syst. Prog., 139-74, California Institute of Technology, Pasadena, CA 91125, USA*
[b] *Div. of Biology, California Institute of Technology, Pasadena, CA 91125, USA*
[c] *Dept. of Electr. Engin., 136-93, California Institute of Technology, Pasadena, CA 91125, USA*

Received 19 December 2003; accepted 29 September 2004
Available online 15 June 2005

Abstract

A key problem in learning representations of multiple objects from unlabeled images is that it is a priori impossible to tell which part of the image corresponds to each individual object, and which part is irrelevant clutter. Distinguishing individual objects in a scene would allow unsupervised learning of multiple objects from unlabeled images. There is psychophysical and neurophysiological evidence that the brain employs visual attention to select relevant parts of the image and to serialize the perception of individual objects. We propose a method for the selection of salient regions likely to contain objects, based on bottom-up visual attention. By comparing the performance of David Lowe's recognition algorithm with and without attention, we demonstrate in our experiments that the proposed approach can enable one-shot learning of multiple objects from complex scenes, and that it can strongly improve learning and recognition performance in the presence of large amounts of clutter.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Bottom-up attention; Saliency; Selective attention; Object recognition; Object-based attention; Learning; Cluttered scenes

* Corresponding author. Fax: +1 626 796 8876.
E-mail address: walther@klab.caltech.edu (D. Walther).
[1] These authors contributed equally to this work.

1077-3142/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2004.09.004

1. Introduction

Object recognition with computer algorithms has seen tremendous progress over the past years, both for specific domains such as face recognition [1–5] and for more general object domains [6–11]. Most of these approaches require segmented and labeled objects for training, or at least that the training object is the dominant part of the training images. None of these algorithms can be trained on unlabeled images that contain large amounts of clutter or multiple objects.

But what is an object? A precise definition of "object," without taking into account the purpose and context, is of course impossible. However, it is clear that we wish to capture the appearance of those lumps of matter to which people tend to assign a name. Examples of distinguishing properties of objects are physical continuity (i.e., an object may be moved around in one piece), having a common cause or origin, having well defined physical limits with respect to the surrounding environment, being made of a well-defined substance. In principle, a single image taken in an unconstrained environment is not sufficient to allow a computer algorithm, or a human being, to decide where an object starts and another object ends. However, a number of cues which are based on the statistics of our everyday's visual world are useful to guide this decision. The fact that objects are mostly opaque and often homogeneous in appearance makes it likely that areas of high contrast (in disparity, texture, color, and brightness) will be associated to their boundaries. Objects that are built by humans are often designed to be easily seen and discriminated from their environment.

Imagine a situation in which you are shown a scene, e.g., a shelf with groceries, and later you are asked to identify which of these items you recognize in a different scene, e.g., in your grocery cart. While this is a common situation in everyday life and easily accomplished by humans, none of the conventional object recognition methods is capable of coping with this situation. How is it that humans can deal with these issues with such apparent ease?

The human visual system is able to reduce the amount of incoming visual data to a small but relevant amount of information for higher-level cognitive processing. Two complementary mechanisms for the selection of individual objects have been proposed, bottom-up selective attention and grouping based on segmentation. While saliency-based attention concentrates on feature *contrasts* [12], grouping and segmentation attempt to find regions that are *homogeneous* in certain features [13,14]. Grouping has been applied successfully to object recognition [15,16]. In this paper, we explore bottom-up attention. In particular, we postulate that a bottom-up attentional mechanism that is designed to respond to areas of high contrast, will frequently select image regions that correspond to objects. Our experiments are designed to test this hypothesis.

Attention is the process of selecting and gating visual information based on saliency in the image itself (bottom-up), and on prior knowledge about scenes, objects and their interrelations (top-down) [17,18]. Upon closer inspection, the "grocery cart problem" (also known as the "bin of parts problem" in the robotics community) poses two complementary challenges—serializing the perception and learning of

Download English Version:

https://daneshyari.com/en/article/9669558

Download Persian Version:

https://daneshyari.com/article/9669558

Daneshyari.com