Computer Vision and Image Understanding 114 (2010) 700–711

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

A framework for visual-context-aware object detection in still images

Roland Perko *, Aleš Leonardis

Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1001 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 9 September 2008
Accepted 22 March 2010
Available online 27 March 2010

Keywords:
Visual context
Object detection
Context integration

ABSTRACT

Visual context provides cues about an object's presence, position and size within the observed scene, which should be used to increase the performance of object detection techniques. However, in computer vision, object detectors typically ignore this information. We therefore present a framework for visual-context-aware object detection. Methods for extracting visual contextual information from still images are proposed, which are then used to calculate a prior for object detection. The concept is based on a sparse coding of contextual features, which are based on geometry and texture. In addition, bottom-up saliency and object co-occurrences are exploited, to define auxiliary visual context. To integrate the individual contextual cues with a local appearance-based object detector, a fully probabilistic framework is established. In contrast to other methods, our integration is based on modeling the underlying conditional probabilities between the different cues, which is done via kernel density estimation. This integration is a crucial part of the framework which is demonstrated within the detailed evaluation. Our method is evaluated using a novel demanding image data set and compared to a state-of-the-art method for context-aware object detection. An in-depth analysis is given discussing the contributions of the individual contextual cues and the limitations of visual context for object detection.

1. Introduction

Objects tend to co-vary with other objects and particular environments, providing a rich collection of contextual associations [29]. It is well known from the literature on visual cognition [31,10,4] and cognitive neuroscience [2,12,1], that the human and animal visual systems use these relationships to improve their ability of categorization. Consequently, context should be used in computer vision as well and can help object detection, as shown in [45,5,18,46,34,29].

In this paper we introduce a framework that uses two different types of visual contextual information to improve object detection. The first type is the spatial relation between an object and its surrounding. It is determined by exploring the visual content of a given scene. The second type are spatial relations between a specific object and other objects in the scene where the context is represented by spatial object co-occurrence (also called object-to-object priming in [16]). Both concepts are visualized in Fig. 1. In the first case semantic information could be extracted from images and exploited to find out if the object is in context, i.e. coherent w.r.t. the scene. In the second case, relative location priors could help to distinguish between correct and incorrect detections, e.g. in urban scenes pedestrians should be more or less at the same height or windows should occur above pedestrians. Overall, we employ the basic ideas as Biederman [3] when he stated that "*object identification is facilitated when an object is presented in a coherent scene*".

1.1. Our contributions

In this work we present a complete framework for visual-context-aware object detection. We answer two major questions which are essential for such a system, i.e. how to represent visual context and how to combine this information with object detection.

First, we propose methods of how to extract visual contextual information from single images and how this information can be learned from examples. For doing so, we utilize a method for sparse coding of contextual features using a spatial sampling technique. Appropriate image features based on geometry and texture are discussed as well. As a result a prior for object detection can be extracted. In addition, we define object co-occurrences and bottom-up saliency as further contextual cues.

Second, we introduce a concept to integrate the contextual information with a local appearance-based object detector. Our mathematical framework and the specific modeling of the conditional probability density functions used for integrating visual context with object detection is a crucial part of this work. As seen later, this modeling contributes significantly to the success of our method.

* Corresponding author. Fax: +386 1 4264 647.
E-mail address: roland.perko@fri.uni-lj.si (R. Perko).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.03.005

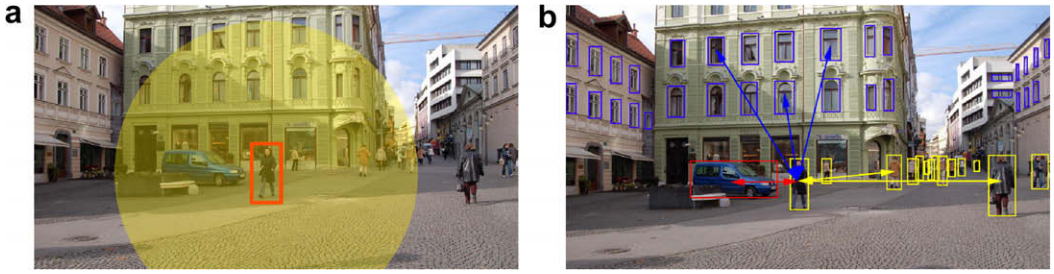

Fig. 1. Definitions of visual contextual information: spatial relations (a) between an object and its surrounding based on the visual content of the scene (e.g. the image content of the given yellow circular region) and (b) between a specific object and other objects in the scene (shown for the object categories pedestrians, cars and windows). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In contrast to other works the proposed integration technique avoids two common mistakes in literature: (i) Researchers assume statistically independence of local-appearance and visual context [47,45,5]. We model these dependency accordingly. (ii) Researches use the output of an object detector as a probability measure [47]. We do not rely on a detector's output, instead we model the probabilities correctly.

For modeling the underlying multi-dimensional probability density functions, we propose to use a *kernel density estimation (KDE)*.

To evaluate the system's performance, we chose the task of pedestrian detection in urban images using a state-of-the-art pedestrian detector on a demanding image database. The evaluation shows that our definition and integration of visual context increases the initial local appearance-based detection rate significantly and outperforms other frameworks for contextual processing. Using the gained insights we discuss the results of visual-context-aware object detection and show its limitations.

1.2. Organization of the paper

Related work will be discussed in detail in Section 2. After that our framework of context-aware object detection and implementation details are described in Section 3, followed by an in-depth analysis of results in Section 4. In Section 5 we discuss the limitations of contextual processing and conclude the paper with Section 6.

2. Related work

The influential work from Oliva and Torralba, e.g. [28,47,44,45], introduced a novel global image representation. An image is decomposed by a bank of multi-scale oriented filters, in particular four scales and eight orientation. The magnitude of each filter is averaged over 16 non-overlapping blocks in a 4×4 grid. The resulting image representation is a 512-dimensional feature vector, which is represented by the first 80 principal components. Despite the low dimensionality of this representation, it preserves most relevant information and is used for scene categorization, such as a landscape or an urban environment. Machine learning provides the relationship between the global scene representation and the typical locations of objects belonging to that category. To the best of our knowledge there exist no evaluation for the combination of this derived context priors with a state-of-the-art object detection algorithm. In a real scenario a coarse prior for the possible object location in the image does not automatically increase the performance of an object detector. As will be seen later, when combined just by multiplication, the results of the object detection may and often do degrade.

Hoiem et al. [17] provided a method to extract the spatial context of a single image. The image is first segmented into so called superpixels, i.e. a set of pixels that have similar properties. These regions are then described by low level image features, i.e. color, texture, shape and geometry, forming a feature vector. Each region is classified into a semantic class, namely *ground*, *vertical structures* and *sky*, using a classifier based on AdaBoost with weak decision tree classifiers. As a result each pixel in the input image is associated with the probabilities of belonging to these three classes. For the task of object detection this classification provides useful cues and they are exploited in [18,34]. Hoiem et al. [18] use the coarse scene geometry to calculate a viewpoint prior and therefore the location of the horizon in the image. The horizon, being the line where the ground plane and the sky intersect in infinity, provides information about the location and sizes of objects on the ground plane, e.g. pedestrians or cars. The scene geometry itself limits the location of objects on the ground plane, e.g. no cars behind the facade of a building. The innovative part of their work is the combination of the contextual information with the object hypotheses using inference. They construct a graphical model of conditional independence for viewpoint, object identities and 3D geometry of surfaces surrounding the objects. The inference is solved using Pearl's belief propagation algorithm [33]. Overall, the main idea is to find the object hypotheses that are consistent in terms of size and location, given the geometry and horizon of the scene. As a result, a cluster of object hypotheses is determined, that fits the data best. This contextual inference uses the global visual context and the relation between objects in that scene. The position of the horizon is an integral part of this system, limiting the approach to object categories that are placed on the ground plane and to objects of approximately the same size. E.g. the approach cannot be used to detect windows on facades or trees.

Bileschi [5] classifies an image into four pre-defined semantic classes. These classes indicate the presence of buildings, roads, skies, and trees, which are identified using their texture properties. These classes are learned from different sets of *standard model features* (also known as HMAX [43]). Bileschi then defines the context using low-level visual features from the *Blobworld* system [6] (three color and three texture-based features). In addition ten absolute image positions are encoded followed by four binary sematic features, representing the four extracted classes (building, road, sky and tree). To extract a context vector for a given position in the image, the data is sampled relative to the object's center for 5 radii and 8 orientations, which results in an 800-dimensional feature vector. However, when using this type of contextual information for object detection, in addition to a standard appearance-based approach, the gain in the detection rate is negligible. This is also confirmed in [53]. A more interesting outcome of the extensive studies by Bileschi is that using global position features (also applied by Torralba and Hoiem) indeed helps to improve the detection rate, due to the input image data being biased. In Bileschi's image database for example, cars are more likely to be in the lower half of the image, because the horizon is in the center of each image.

There are a couple of additional works on visual context, however, they are only sparsely related to our work. For example the