Image and Vision Computing 29 (2011) 569–579

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

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

Toward coherent object detection and scene layout understanding ☆

Sid Yingze Bao *, Min Sun, Silvio Savarese

Univsersity of Michigan at Ann Arbor, Ann Arbor, MI, 48105, USA

ARTICLE INFO

Article history:
Received 3 May 2011
Received in revised form 25 July 2011
Accepted 4 August 2011

Keywords:
Object detection
Scene layout
Focal length estimation
Supporting surface estimation

ABSTRACT

Detecting objects in complex scenes while recovering the scene layout is a critical functionality in many vision-based applications. In this work, we advocate the importance of geometric contextual reasoning for object recognition. We start from the intuition that objects' location and pose in the 3D space are not arbitrarily distributed but rather constrained by the fact that objects must lie on one or multiple supporting surfaces. We model such supporting surfaces by means of hidden parameters (i.e. not explicitly observed) and formulate the problem of joint scene reconstruction and object recognition as the one of finding the set of parameters that maximizes the joint probability of having a number of detected objects on K supporting planes given the observations. As a key ingredient for solving this optimization problem, we have demonstrated a novel relationship between object location and pose in the image, and the scene layout parameters (i.e. normal of one or more supporting planes in 3D and camera pose, location and focal length). Using a novel probabilistic formulation and the above relationship our method has the unique ability to jointly: i) reduce false alarm and false negative object detection rate; ii) recover object location and supporting planes within the 3D camera reference system; iii) infer camera parameters (view point and the focal length) from just one single uncalibrated image. Quantitative and qualitative experimental evaluation on two datasets (desk-top dataset [1] and LabelMe [2]) demonstrates our theoretical claims.

1. Introduction

When we observe a complex scene such as an office or a street, it is easy for our visual system to recognize the objects and infer their spatial organization in the environment. Objects do not appear in arbitrary locations: it is very unlikely to observe a monitor floating in the air or a car hanging from a building. Objects are rather organized in the physical space in consistent geometrical configurations: their locations and poses obey the law of physics (objects lie on supporting planes in stable configurations) and follow the conventions of human behavior. It is clear that when humans observe the environment, such constraints will help reinforce the process of joint recognition and scene layout recovery [3,4]. The recognition of objects with the estimate of their locations, scales and poses helps infer the spatial properties of the environment (e.g., the location and orientation of the surface where objects lie), and in turn the scene layout provides strong spatial contextual cues as for where and how objects are expected to be found. Contributions in computer vision for the past decade have followed the common paradigm of recognizing objects in isolation [5–9], regardless of the geometrical contextual cues. It is indeed true that objects can be in general recognized even when no information about the scene layout is provided. However, we claim that joint object recognition and scene reconstruction are critical if one wants to obtain a coherent understanding of the scene as well as minimize the risk of detecting false positive examples or missing true positive ones. This ability is crucial for enabling higher level visual tasks such as event or activity recognition and in applications such as robotics, autonomous navigation, and video surveillance.

The intuition that recognition and reconstruction are mutually beneficial has been initially explored in early works of computer vision [10–15], and recently revitalized in [16–27]. In Hoiem et al. [16], the process of detecting objects in a complex scene is enhanced by introducing the geometrical contextual information of the scene layout [28] (e.g., vertical surfaces, ground horizontal planes, etc.). More explicit reasoning on the relationship between supporting planes and objects hasbeen investigated in Hoiem et al. [29] and Hedau et al. [17,18]. Hedau et al. [17,18] introduced a flexible methodology for estimating the layout of indoor scenes by modeling the scene using a3D cube representation. Following our preliminary study [30], we too advocate the importance of geometrical contextual reasoning for object recognition and focus on demonstrating that the contextual cues provided by object location and pose can be used, in turn, to reinforce the detection and prune out false alarms (Fig. 1). Our key idea is that objects' locations and poses in the 3D space are not arbitrarily distributed but rather constrained by the fact that objects must lie on one or multiple supporting surfaces. We model such

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author.
E-mail addresses: yingze@umich.edu (S.Y. Bao), sunmin@umich.edu (M. Sun), silvio@eecs.umich.edu (S. Savarese).
URL: http://www.eecs.umich.edu/~yingze (S.Y. Bao).

0262-8856/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2011.08.001

supporting surfaces by hidden parameters (i.e. not explicitly observed) and seek a configuration of objects and supporting surfaces in the 3D space that best explains the observations, including the estimation of each object's location, scale and pose. To this end, we leverage on recent methods for detecting multi-category objects and estimating their poses accurately from a single image [31–36]. Unlike [16], where contextual information was partially provided by the explicit estimation of surface orientation using the geometric context operator [28], we only use the objects *themselves* for extracting contextual cues. Thus, we do not require supporting planes or other scene surfaces to be visible or detectable in order to perform the joint recognition and reconstruction. Rather, supporting planes are implicitly estimated from the observation of the object location and pose in the image. Moreover, our camera representation is general: We only hypothesize that the camera has zero skew and unit pixel ratio (but unknown focal length). Most importantly, we do not make the assumption that the camera is at fixed distance from the ground plane and has a fixed view angle. Because of these properties, our algorithm can be successfully applied in both outdoors and indoors scenarios. Notice that Hedau etal. [17,18] use cues such as vanishing lines that are complementary to ours and could be nicely integrated into our framework. Also notice that physics-based constraints such as those introduced in Gupta et al. [26] enable different ways for modeling the interaction between scene and objects wherein, in this case, objects are mostly identified as urban elements (i.e., buildings and houses). Finally, in Payet et al. [27] the analysis of textures is introduced to provide scene-specific constraints among objects.

The main contributions of our work include: 1. A novel representation that can jointly model 3D objects locations and 3D supporting surfaces (planes) from the observations in a single image. Concretely, the problem of joint scene reconstruction and object recognition is formulated as finding a set of parameters that maximize the joint probability of having a number of detected objects on K supporting planes given the observations (Section 2). 2. A relationship that connects the 2D image observation of object location and zenith angle pose with the normals of the supporting planes and with the camera focal length parameter. We prove that this relationship yields necessary conditions for estimating the camera focal length and the supporting planes' 3D orientations and locations (in the camera reference system) from the locations and zenith poses of at least 3 objects in the image. The relationship is general in that objectsdo not necessarily need to lie on the *same* supporting plane as long as their supporting planes are parallel with respect to each other and the objects are not collinear (Section 3.1). 3. A framework that uses the above relationships and a novel probabilistic formulation to jointly detect objects (so as to reduce false alarm and false negative rates) and recover (within the camera reference system) the objects' 3D locations, the 3D supporting planes, and the camera focal length parameter. All of the outcomes mentioned above are merely based on one single semi-calibrated image (Section 2). Experimental evaluation on two datasets (desk-top dataset [1] and the car and pedestrian Label-Me dataset [2]) demonstrates our theoretical claims (Section 4).

2. Modeling objects and scene layout

Given an image portraying a number of objects, our work proposes a new model for jointly recognizing objects in the scene and recovering the scene layout that best "explains" the evidence measured in the image. In this paper, the term "scene layout" indicates: i) the objects' 3D locations and poses in the camera reference system; ii) the 3D locationand orientation of their supporting planes in the camera reference system; iii) the camera focal length. In this section we will first introduce notations and assumptions and then formulate the problem.

2.1. Assumptions and notations

We assume that each object lies on a supporting plane at an up-right pose. This assumption is satisfied in most real world scenes. For example, a car is usually touching the ground by four wheels rather than only two and a wineglass is usually standing vertically rather than obliquely (Fig. 2). This is consistent with the common observation that objects rarely float in the air or appear in unstable poses. Furthermore, if multiple supporting planes co-exist in one image, we assume that these planes are all parallel to each other. This parallel relationship of planes holds for most daily-life scenes. Notice that we are *not* assuming *the camera* must be "up-right" with respect to the supporting surfaces.

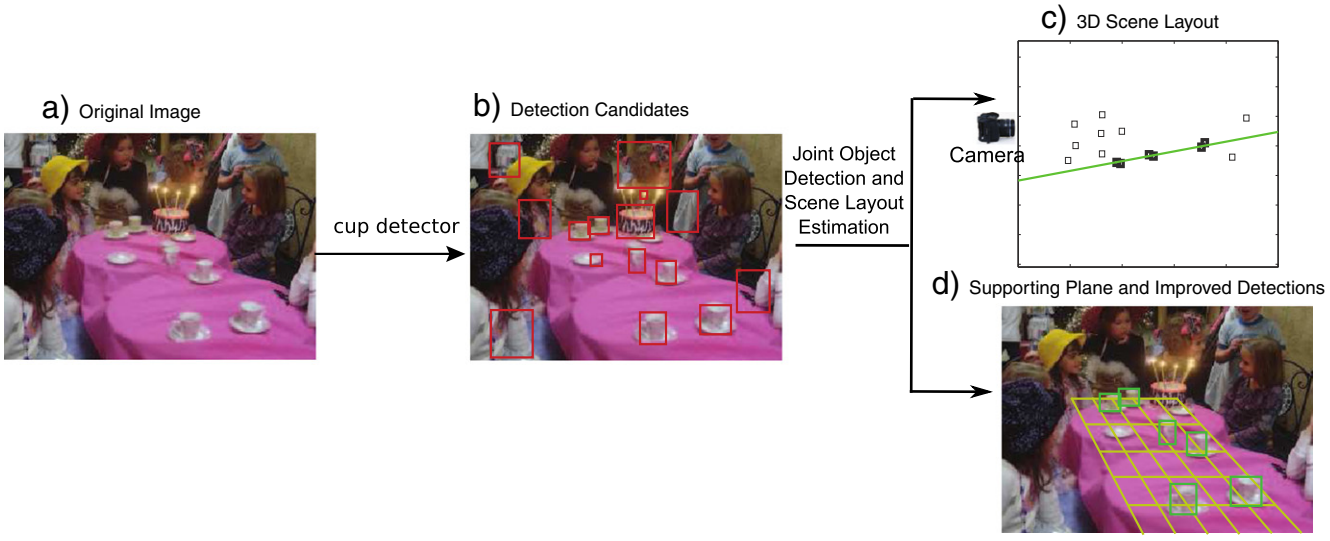

Fig. 1. A conceptual illustration of the flowchart of our algorithm. (a) Original input image with unknown camera parameters; (b) Detection candidates provided by a baseline "cup" detector; (c) The 3D layout. The figure shows the side view of the 3D reconstructed scene. The supporting plane is shown in green. Dark squares indicate the objects detected and recovered by our algorithm; light squares indicate objects detected by the baseline detector and identified as false alarms by our algorithm; (d) Our algorithm detects objects and recovers object locations and supporting plane (in gold color) orientations and locations within the 3D camera reference system from one single image. We show only a portion of the recovered supporting plane for visualization purposes.