Computer Vision and Image Understanding 117 (2013) 1190–1202

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

Object detection, shape recovery, and 3D modelling by depth-encoded hough voting ☆

Min Sun [a,*], Shyam Sunder Kumar [a], Gary Bradski [b], Silvio Savarese [a]

[a] Department of Electrical and Computer Engineering, The University of Michigan, Ann Arbor, MI 48105, United States
[b] Founder/Chief Scientist at Industrial Perception Inc., CA, United States

ARTICLE INFO

Article history:
Received 24 July 2012
Accepted 10 May 2013
Available online 22 May 2013

Keywords:
Object recognition
Object detection
Viewpoint estimation
Shape recovery
3D reconstruction
Shape completion
Texture completion

ABSTRACT

Detecting objects, estimating their pose, and recovering their 3D shape are critical problems in many vision and robotics applications. This paper addresses the above needs using a two stages approach. In the first stage, we propose a new method called DEHV – Depth-Encoded Hough Voting. DEHV jointly detects objects, infers their categories, estimates their pose, and infers/decodes objects depth maps from either a single image (when no depth maps are available in testing) or a single image augmented with depth map (when this is available in testing). Inspired by the Hough voting scheme introduced in [1], DEHV incorporates depth information into the process of learning distributions of image features (patches) representing an object category. DEHV takes advantage of the interplay between the scale of each object patch in the image and its distance (depth) from the corresponding physical patch attached to the 3D object. Once the depth map is given, a full reconstruction is achieved in a second (3D modelling) stage, where modified or state-of-the-art 3D shape and texture completion techniques are used to recover the complete 3D model. Extensive quantitative and qualitative experimental analysis on existing datasets [2–4] and a newly proposed 3D table-top object category dataset shows that our DEHV scheme obtains competitive detection and pose estimation results. Finally, the quality of 3D modelling in terms of both shape completion and texture completion is evaluated on a 3D modelling dataset containing both in-door and out-door object categories. We demonstrate that our overall algorithm can obtain convincing 3D shape reconstruction from just one single uncalibrated image.

Published by Elsevier Inc.

1. Introduction

Detecting objects and estimating their geometric properties are crucial problems in many application domains such as robotics, autonomous navigation, high-level visual scene understanding, surveillance, gaming, object modelling, and augmented reality. For instance, if one wants to design a robotic system for grasping and manipulating objects, it is of paramount importance to encode the ability to accurately estimate object orientation (pose) from the camera view point as well as recover structural properties such as its 3D shape. This information will help the robotic arm grasp the object at the right location and successfully interact with it. Moreover, if one wants to augment the observation of an environment with virtual objects, the ability to reconstruct visually pleasing 3D models for object categories is very important.

This paper addresses the above needs, and tackles the following challenges: (i) Learn models of object categories by combining view specific depth maps along with the associated 2D image of object instances of the same class from different vantage points. Depth maps with registered RGB images can be easily collected using sensors such as Kinect Sensor [5]. We demonstrate that combining imagery with 3D information helps build richer models of object categories that can in turn make detection and pose estimation more accurate. (ii) Design a coherent and principled scheme for detecting objects and estimating their pose from either just a single image (when no depth maps are available in testing) (Fig. 1b), or a single image augmented with depth maps (when these are available in testing). In the latter case, 3D information can be conveniently used by the detection scheme to make detection and pose estimation more robust than in the single image case. (iii) Have our detection scheme reconstruct the 3D model of the object from just a single uncalibrated image (when no 3D depth maps are available in testing) (Fig. 1c–g) and without having seen the object instance during training.

In this paper, we propose a two stages approach to address the above challenges (Fig. 2). In the first stage, our approach seeks to (i) detect the object in the image, (ii) estimate its pose, and (iii) recover a rough estimate of the object 3D structure (if no depth maps are available in testing). This is achieved by introducing a new

☆ This paper has been recommended for acceptance by Carlo Colombo.

* Corresponding author.
E-mail addresses: sunmin@umich.edu (M. Sun), shyamsk@umich.edu (S.S. Kumar), garybradski@gmail.com (G. Bradski), silvio@eecs.umich.edu (S. Savarese).

1077-3142/$ - see front matter Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.cviu.2013.05.002

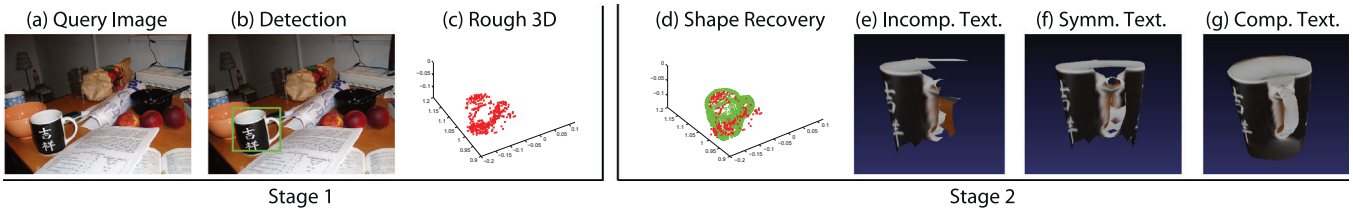

Fig. 1. Key steps of our reconstruction algorithm: (a) Single query 2D image. (b) Detected object; the bounding box indicates the location where the object has been estimated in the image; Our proposed Depth Encoded Hough Voting (DEHV) detector can be used to recognize object class label, roughly estimate the object pose (i.e., object orientation in the camera reference system), and automatically reconstructs surface elements (3D points) in the camera reference system (c). As figure shows, the reconstruction is clearly partial and incomplete; (d) Shape recovery: by using the estimated object class label and pose, we propose a novel 2D + 3D ICP algorithm to register the reconstructed surface elements with one of the 3D models that is available in training; this allows to infer the object 3D structure in regions that are not visible from the query image. (e) Texture mapping: after performing 3D shape registration, we texture map image texture to the 3D shape model; again, the object texture is incomplete as we cannot map image texture to occluded surface elements. (f) Texture completion: we use the fact that some object categories are symmetric to transfer image texture to the occluded regions. (g) Remaining un-textured surfaces elements are completed using image compositing methods inspired by [6].

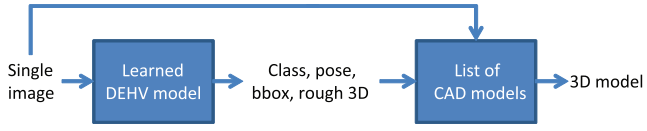

Fig. 2. Flow chart showing the process of our proposed system.

formulation of the Implicit Shape Model (ISM) [1] and generalized Hough voting scheme [7]. In our formulation, depth information is incorporated into the process of learning distributions of object image patches that are compatible with the underlying object location (shape) in the image plane. We call our scheme *DEHV – Depth-Encoded Hough Voting scheme* (Section 3.1). DEHV addresses the intrinsic weaknesses of existing Hough voting schemes [1,8–10] where errors in estimating the scale of each image object patch directly affects the ability of the algorithm to cast consistent votes for the object existence. To resolve this ambiguity, we take advantage of the interplay between the scale of each object patch in the image and its distance (depth) from the corresponding physical patch attached to the 3D object, and specifically use the fact that objects (or object parts) that are closer to the camera result in image patches with larger scales. Depth is encoded in training by using available depth maps of the object from a number of view points. At recognition time, DEHV is applied to detect objects (Fig. 1b), estimate their pose, and simultaneously infer their 3D structure given hypotheses of detected objects (Fig. 1c). The object 3D structure is inferred at recognition time by estimating (decoding) the depth (distance) of each image patch involved in the voting from the camera center. Critically, depth decoding can be achieved even if just a single test image is provided. If depth maps are available in testing, the additional information can be used to further validate if a given detection hypothesis is correct or not. We summarize the inferred quantities in Fig. 3 and the required supervision in Fig. 4. Notice that the inferred object 3D structure

Single Image		
	Depth in testing	No depth in testing
Inferred quantities	object class, location, scale, pose	object class, location, scale, pose, depth map

Fig. 3. Estimated quantities in Stage 1.

Stage 1	Stage 2
-Images of object from multiple views -Depth maps of object from multiple views -Bounding boxes and pose annotation	-List of CAD Models

Fig. 4. Required degree of supervision in training for each stage.

from stage one is partial (it does not account for the portions of the object that are not visible from the query image) and sparse (it only recovers depth for each voting patch). The goal of the second stage is to obtain a full 3D object model where both 3D structure and albedo properties (texture) are also recovered.

In the second stage, the information inferred from stage one (object location in the image, scale, pose, and rough 3D structure) is used to obtain a full 3D model of the object. Specifically, we consider a 3D modelling stage where a full 3D model of the object is obtained by 3D shape recovery and texture completion (Section 3.2). We carry out 3D shape recovery (i.e., infer shape from the unseen regions) by: (i) utilizing 3D shape exemplars from a database of 3D CAD models which can be collected from [11] and other online 3D warehouses, or obtained by shape from silhouette [12] and (ii) applying a novel 2D + 3D iterative closest point (ICP) matching algorithm which jointly registers the best 3D CAD model to the inferred 3D shape and the occlusion boundaries of back projected 3D CAD model to object contours in the image. By choosing the best fit, our system obtains a plausible full reconstruction of the object 3D shape (Section 3.3) (Fig. 1d). Object appearance is rendered by texture mapping the object image into the 3D shape. Such texture is clearly incomplete as non-visible object surface areas cannot be texture mapped (Fig. 1e). Thus, we perform texture completion by: (i) transferring texture to such non-visible object surface areas by taking advantage of the fact that some object categories are symmetric (when possible) (Fig. 1f) and (ii) using an error-tolerant image compositing technique inspired by [6] to fill the un-textured regions (i.e., holes) (Section 3.4) (Fig. 1g). We summarize the required supervision in Fig. 4.

Extensive experimental analysis on a number of public datasets (including car Pascal VOC07 [2], mug ETHZ Shape [3], mouse and stapler 3D object dataset [13]), an two in-house datasets (comprising at most five object categories), where ground truth 3D information is available, are used to validate our claims (Section 4). Experiments with the in-house datasets demonstrate that our