Expert Systems with Applications 42 (2015) 8123–8133

Contents lists available at ScienceDirect

Expert Systems with Applications

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

What, Where and How? Introducing pose manifolds for industrial object manipulation

R. Kouskouridas [a,*], A. Amanatiadis [b], S.A. Chatzichristofis [c], A. Gasteratos [b]

[a] Department of Electrical & Electronic Engineering, Imperial College, South Kensington Campus, SW7 2AZ London, UK
[b] Department of Production & Management Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
[c] Department of Electrical & Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece

ARTICLE INFO

Article history:
Available online 3 July 2015

Keywords:
Object grasping
Object recognition
Pose estimation
Ontology-based semantic categorization

ABSTRACT

In this paper we propose a novel method for object grasping that aims to unify robot vision techniques for efficiently accomplishing the demanding task of autonomous object manipulation. Through ontological concepts, we establish three mutually complementary processes that lead to an integrated grasping system able to answer conjunctive queries such as *"What"*, *"Where"* and *"How"*? For each query, the appropriate module provides the necessary output based on ontological formalities. The *"What"* is handled by a state of the art object recognition framework. A novel 6 DoF object pose estimation technique, which entails a bunch-based architecture and a manifold modeling method, answers the *"Where"*. Last, *"How"* is addressed by an ontology-based semantic categorization enabling the sufficient mapping between visual stimuli and motor commands.

1. Introduction

Contemporary vision-based robotic systems tackle the object manipulation problem by extracting appearance features that are to be matched with the ones already contained in the training dataset (Wang, Tao, Di, Ye, & Shi, 2012). However, these systems fail to generalize to objects not included in the training set, whilst they are highly depended on the architecture of the respective robotic platform. It is apparent that a beyond the state of the art methods for automatic object grasping, e.g. targets placed on a conveyor belt, should: (i) be capable of manipulating any object offering large generalization capacities; (ii) be based on low dimensional input vectors, thus, resulting to minimum system complexity; (iii) execute in real-time and (iv) be invariant to the robot's architecture (Da Xu, Wang, Bi, & Yu, 2012).

Similar to any other robotic task, the human hand-gripping outperfomers any robotic grasping system and remains the ultimate standard. The brain and hand are the two primary determinants of the human grasping action and attempting to separately imitate each of them when trying to reproduce this polymodal task proves to be insufficient. Consequently, any interaction between them in terms of knowledge requirements and reasoning capabilities should be sought (Liu, 2011). The problem of shape extraction with non discriminative local features for object grasping was analyzed in Ying, Fu, and Pollard (2007), by synthesizing humanlike enveloping grasps and utilizing a shape matching algorithm. Such approaches attempt to answer certain questions based on the different constraints, e.g. one might possess specific knowledge of where the graspable part is, yet the question of how to grasp it remains. In fact, trying to answer solely each of the three questions, namely *What*, *Where*, and *How*, leaves out critical semantic constraints that affect the whole context of the object grasping action. Even for tasks where the object to be grasped is known, depending on the operational scenario, different semantic constraints are introduced. The latter determine the way the object will be grasped according to the affordances and the attributes the specific task exhibits. For example, the way a pencil is held is different for writing than for sharpening it. Hence, the question *"what is the object to be grasped?"* is not sufficient to complete the action, but the answer depends also on how exactly the object is expected to be used (Bicchi, 2000).

Bin-picking stands for one of the most widely encountered industrial applications where robots are asked to automatically manipulate similar objects usually placed in bins or boxes. Severe occlusions, foreground clutter and large scale changes are among the cascading issues that put additional barriers to this challenging problem. Liu et al. (2012) presented a chamfer matching-based solution that extract depth edges via a multi-flash camera, while Sansoni, Bellandi, Leoni, and Docchio

* Corresponding author.
E-mail addresses: r.kouskouridas@imperial.ac.uk (R. Kouskouridas), aamanat@ee.duth.gr (A. Amanatiadis), schatzic@ee.duth.gr (S.A. Chatzichristofis), agaster@pme.duth.gr (A. Gasteratos).

http://dx.doi.org/10.1016/j.eswa.2015.06.039
0957-4174/

(2014) showed how a laser source scanning architecture can facilitate accurate pose estimation. In Buchholz, Kubus, Weidauer, Scholz, and Wahl (2014) inertial and visual data are fused to calculate grasp poses of testing objects (Kuo, Su, Lai, & Wu, 2014). In turn, in Nieuwenhuisen et al. (2013) and Buchholz, Futterlieb, Winkelbach, and Wahl (2013) 3D descriptors (shape-based and spin images, respectively) are extracted from RGB-D input data and fed to nearest-neighbor classifiers to acquire accurate recognition and pose estimation results.

In this paper, we aim at providing a consolidated architecture for automatic grasping tasks, which can provide answers to the next questions: "*What is the item?*", "*Where is the item placed?*" and "*How can I manipulate it?*". Thereupon, we assess a shape-based methodology for the recognition task and we acquire exact detection results via a Bag-of-Features classification procedure. In addition, the pose estimation module relies on the notion that even unlike objects when perceived under similar perspectives should hold respective similar poses. Grasping points are determined by means of an ontology, where the recognized objects inherit accurate grasping coordinates from the relevant class. The proposed ontology includes: (i) object-class associated data, (ii) a pose manifold for each instance of the object-class conceptual model and (iii) the grasping points information of any trained instance. The basic concept of this procedure is depicted in Fig. 1.

Our main contributions can be summarized as follows: Compared to the state of the art works in object recognition and pose estimation (Brachmann et al., 2014; Bonde, Badrinarayanan, & Cipolla, 2014; Hinterstoisser et al., 2011; Lim, Khosla, & Torralba, 2014; Tejani, Tang, Kouskouridas, & Kim, 2014; Wohlhart & Lepetit, 2015) our method offers higher generalization capabilities through the recognition of objects that do not have to belong in the training dataset. Additionally, our sophisticated manifold modeling technique builds compact and object-class invariant manifolds that are not prone to occlusions. Moreover, the paper in hand represents the first integrated research attempt in industrial-centric ontologization focusing on the liaison between

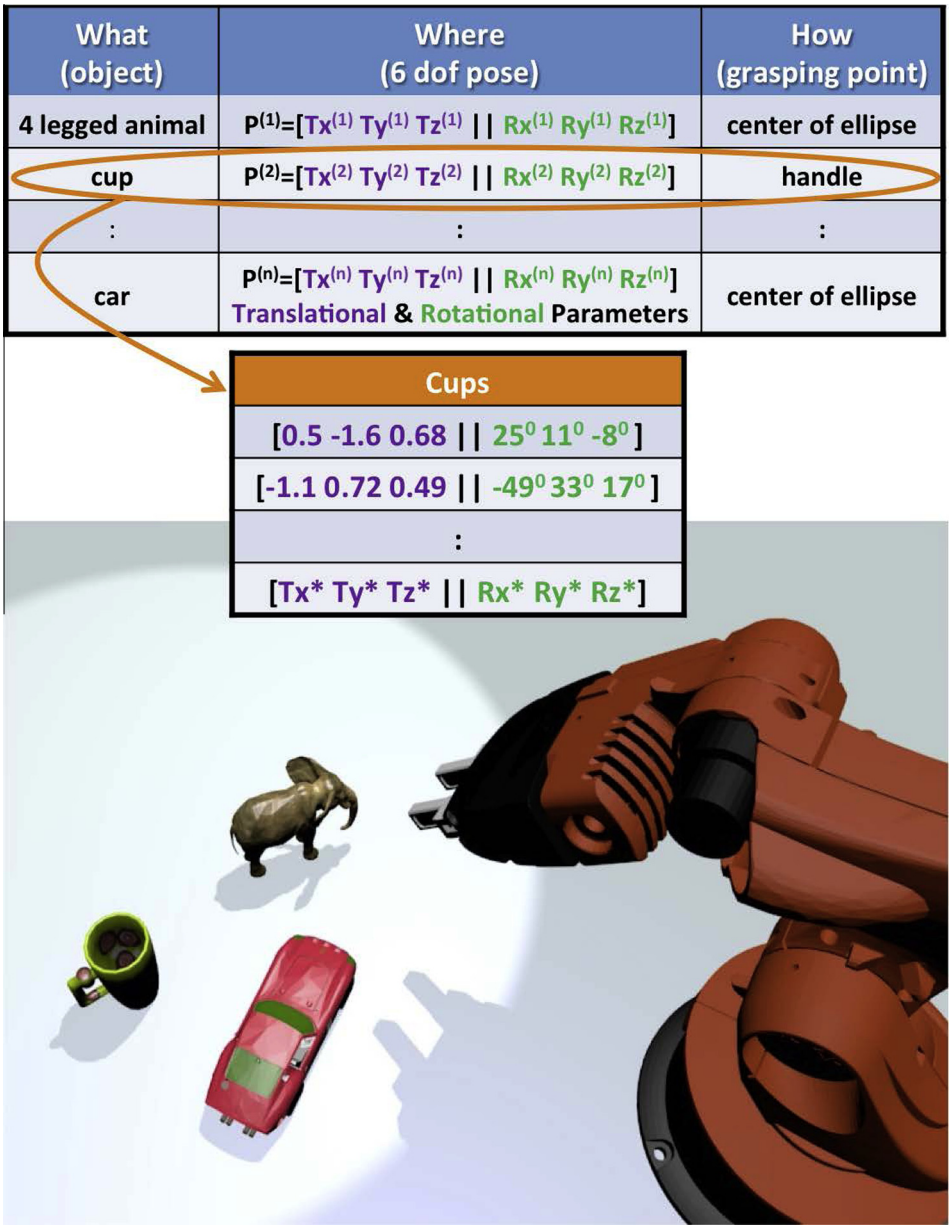

Fig. 1. The proposed architecture aims at providing an efficient solution to the autonomous unknown object manipulation problem by addressing the challenging issues risen during the recognition, pose estimation and grasping point calculation tasks.