2015

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

Kouskouridas, R.

Elsevier

Expert Systems with Applications xxx (2015) xxx–xxx

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 xxxx

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/© 2015 Published by Elsevier Ltd.

(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

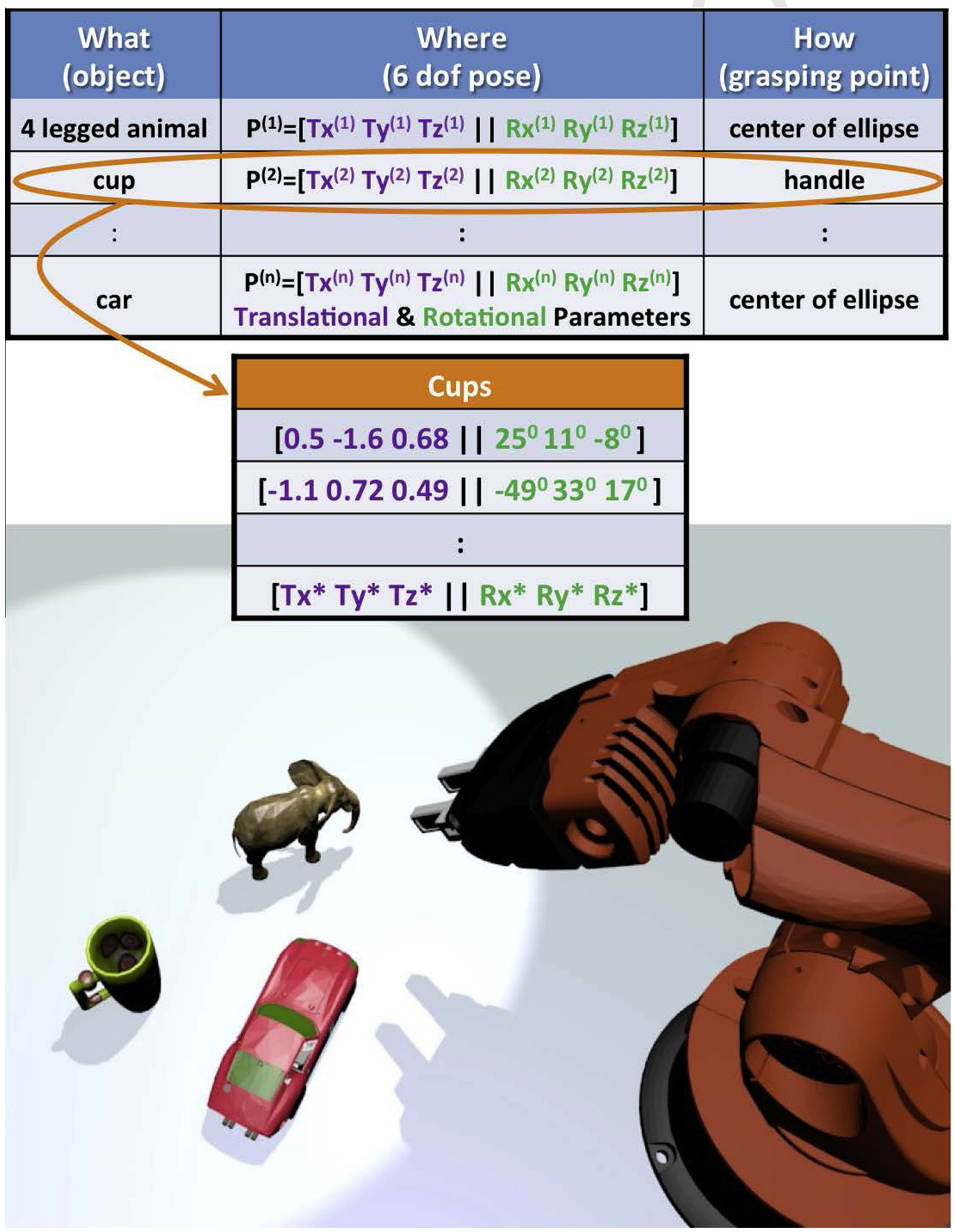

Fig. 1. The proposed architecture aims at providing an efficient solution to the autonomous unknown object manipulation problem by addressing the challenging issues risen during the recognition, pose estimation and grasping point calculation tasks.

Please cite this article in press as: Kouskouridas, R., et al. What, Where and How? Introducing pose manifolds for industrial object manipulation. *Expert Systems with Applications* (2015), http://dx.doi.org/10.1016/j.eswa.2015.06.039

image understanding algorithms and the corresponding motor commands in the particular task of unknown object manipulation. Despite the continuous research developments on ontology-based frameworks for image retrieval, web indexing or even robotics, very limited activity in industrial object manipulation is discerned. Additionally, our method is invariant to the robotic architecture or the distinctive parts used, whilst exhibiting real-time performance. Moreover, the proposed system can be easily adopted and expanded with view to manipulate any variety of objects belonging to different classes without additional training on new targets.

The rest of the paper is organized as follows. In Section 2, we discuss the related work on the three separate cores of our architecture. In Section 3, we demonstrate the methodology of exploring and answering the three primary constraints introduced. In Section 4, we exhibit the experimental results and compare the performance of the proposed framework with other widely used grasping systems through qualitative measures. Finally, we draw concluding remarks in section 5.

2. Related work

Sensorimotor architectures for object grasping try to address the challenges risen by making significant progress in several layers of abstraction (Bannat et al., 2011). While different architectures and systems have been proposed, the main core systems are common; improvements are made in either the core systems or their reciprocal engagements (McGuire et al., 2002; Wang, Ren, Mills, & Cleghorn, 2010). The next subsections present the related work based on the highest layers of our core system, with special emphasis on their mutual interactions.

2.1. Object recognition using content based image retrieval techniques

In the past, content based image retrieval (CBIR) techniques have been adopted in robot grasping systems to facilitate object recognition (Kragic & Christensen, 2003; Steil, Röthling, Haschke, & Ritter, 2004) while they are distinguished into two categories, depending on whether they employ global features (GFs) or local ones (LFs). GFs, are the ones describing the content of an image in a holistic manner and the information described by them concerns either the color, the shape, or the texture of an image (Manjunath, Ohm, Vasudevan, & Yamada, 2001). Despite the fact that in applied research, image retrieval often relies on global features, at least as a foundation for further research (Chatzistavros, Chatzichristofis, Zagoris, & Stamatelos, 2015), they often lead to a query sensitive holistic description of the visual information. In other words, image retrieval using global features is notoriously noisy for image queries of low generality, i.e. the fraction of relevant images in a collection. Image retrieval methods employing global features typically rank the entire collection using some distance measure. Revisiting the example from Arampatzis, Zagoris, and Chatzichristofis (2013) and Papadopoulos, Kalogeiton, Chatzichristofis, and Papamarkos (2013), a query image of a red tomato on white background would retrieve images from the collection that illustrate e.g. a red pie-chart on white paper. In other words, if the collection does not contain visually similar to the query images, early rank positions may be dominated by spurious results such as the pie-chart, which may even be ranked before tomato images on non-white backgrounds. In conclusion, global features are able to retrieve only images with similar visual properties in a holistic way.

On the other hand, retrieval systems which employ LFs, extract the content of an image on a set of 'Points of Interest (POI)', each of which is described using a feature vector invariant in scaling and rotation. The replacement of GFs by LFs, slightly improves the retrieval effectiveness when searching for images with similar visual and conceptual content (Aly, Welinder, Munich, & Perona, 2009; Iakovidou, Anagnostopoulos, Kapoutsis, Boutalis, & Chatzichristofis, 2014). Additionally they equip the respective systems with the capability of identifying objects in cases of occlusions or cluttered backgrounds. Yet, the problem of a system based on LFs is its computational burden. Hence, modern approaches combines LFs' effectiveness with GFs' efficiency. Such an approach is the Bag-of-Features (BoF) -or Bag-of-Visual-Words-model, which originates from the well-known Bag-of-Words paradigm and is regarded as a "*promising framework for CBIR*" (Ren, Collomosse, & Jose, 2011). This model has also been applied in other robotic applications (Kostavelis & Gasteratos, 2013), mostly due to: (i) its better retrieval effectiveness over GF representations and (ii) its better efficiency than LF representations. In the proposed method, the object is classified under one of the classes used to train our system. We use the BoF model to classify an object captured by a single digital camera in one of the predefined classes, by adopting characteristics from the method proposed in Chatzichristofis, Iakovidou, Boutalis, and Marques (2013). The object is captured at angle γ and distance d (the distance between the center of the camera and the object's centroid), both of which are neither constant nor predefined. Thus, the system is expected to identify a 3D object by a 2D projection of it.

2.2. Pose estimation

The adequate implementation of robotic manipulation tasks necessitates the accurate estimation of the 6 DoF pose of the testing object (Kouskouridas, Amanatiadis, & Gasteratos, 2011; Kouskouridas, Charalampous, & Gasteratos, 2014; Popovic et al., 2010; Sansoni et al., 2014). The simplicity along with facile training sessions render template matching methods as one of the most widely used solutions for object detection tasks (Ferrari, Tuytelaars, & Van Gool, 2006; Hinterstoisser et al., 2011; Ma, Chung, & Burdick, 2011; Rios-Cabrera & Tuytelaars, 2013; Tejani et al., 2014). However, the main drawbacks of such techniques are their sensitivity to occlusions and the respective laborious training sessions. Point-to-Point techniques build object models as pairs of points extracted on point clouds (Drost, Ulrich, Navab, & Ilic, 2010). More recently, Brachmann et al. (2014) introduced a new representation in form of a joint 3D object coordinate and class labeling, which, however, suffers in cases of occlusions. Song and Xiao (2014) proposed a computationally expensive approach to the 6 DoF pose estimation problem that slides exemplar SVMs in the 3D space, while in Bonde et al. (2014) shape priors are learnt by soft labeling random forest for 3D object classification and pose estimation. In turn, part-based approaches focus on learning distinctive object models from wide training collections to strive the partial occlusion challenge. Constellation architectures (Cao, Ning, Yan, & Li, 2012) are regarded as an extension of part-based ones since they apply similar strategies to connect distinguishable areas of the object. Although plenty of solutions for object registration exist, to the best of our knowledge, there is hardly any algorithm combining sufficient robustness and low computation load.

In this paper, the 3D object pose estimation is based on a custom manifold modeling technique by means of ellipse fitting. We consider our technique as a mixture of template matching and part-based approach. A similar study (Hinterstoisser, Benhimane, & Navab, 2007) suggests that the 3D pose of an object can be recovered through the extraction of 4 or 5 neighboring primary points with equal distribution over the object's surface. However, this approach results to non-compact and occlusion biased pose models. Another close work, the statistical manifold modeling of Mei, Liu, Hero, and Savarese (2011), which is considered to be a benchmark in manifold fitting, enables accurate registration of objects in their 3D environment. However, the learnt manifolds are based on two additional operations to address intra-class minimization and inter-pose maximization. Moreover, this work makes use of a limited training dataset, thus restricting the pose recovery to only one class.

2.3. Ontologies

The process of linking knowledge derived from complex images to specific primitives with semantic meaning forms an intriguing research topic. From medical image annotation (Hu, Dasmahapatra, Lewis, & Shadbolt, 2003) to image retrieval and classification (Mezaris, Kompatsiaris, & Strintzis, 2003), ontological frameworks provided assistance in machine-based reasoning of the acquired data. In computer science, ontologies, as introduced in Gruber (1993), aim at adding semantics with a view to specify the meanings of annotations. Essentially, an ontology represents a data-driven model representing both the underlying framework and the individual instances along with their definitions of a particular domain. In the field of computer vision, ontologies are not yet mature enough and they are adopted primarily for image retrieval tasks and object classification (Chen, Li, & Kwok, 2011), whilst in the particular task of object manipulation ontologies were employed to allow an efficient object classification along with the respective grasping points (Kouskouridas, Retzepi, Charalampoglou, & Gasteratos, 2012; Vorobieva, Soury, Hède, Leroux, & Morignot, 2010). However, both the aforementioned works fail to generalize to unknown (untrained) objects, whilst requiring adequate knowledge of the working environment of the robot. In other robotics applications ontologies are utilized with view to provide a more compact representation of the 3D objects (Varadarajan & Vincze, 2012) and to study the relation between specific models and the corresponding robot action (Modayil & Kuipers, 2007). In this paper, the ontologies are utilized in an holistic manner with aim to establish a novel knowledge domain focusing on industrial object manipulation.

3. Methodology

The principal concepts of the proposed method are illustrated in the block diagram of Fig. 2. Initially, we collect images of objects contained in large databases dedicated to shape classification and 3D pose recovery. The generalization potential of the method is boosted by accumulating sufficient images of various objects, along with the respective shape silhouettes captured from varying viewpoints. We aim at building ontological concepts able to assist grasping by facilitating "*what*"; "*where*" and "*how*". Thereupon, we employ: (i) an object recognition module, answering to "*what*"; (ii) a 6 DoF object pose estimation technique, replying to "*where*" and (iii) a grasping point calculation algorithm, solving the "*how*". We accumulate the outcome of the aforementioned individual modules in an ontology, in which each recognized object is accompanied by a 3D pose measurement and a set of grasping points.

3.1. Object recognition – What?

Our object recognition module aims at producing accurate identification results, while its underlying idea mimics the properties of the BoF model. The latter suggests dividing the whole procedure into two discriminative phases, viz. the training and the retrieval one. During the training phase, LFs are extracted from the database images. Let R, represent a set of randomly selected features that are classified into m classes, using a well established classifier. The center of each cluster represents a visual word, while the total set of the words – classes in our approach – define the "codebook".

In turn, during the retrieval phase, LFs extracted from each single image are,then classified as per the classes generated during training. We adopt a soft-labeling architecture that allows each of the extracted LFs to be classified in more than one class. By the end of this procedure, each image is represented by a vector of m positions, each of which includes the number of LFs belonging to the class. We equip our recognition module with beyond the state of the art properties by adopting and enriching the method proposed in Chatzichristofis et al. (2013). Towards this end, a number of R features are randomly selected from the database and forwarded into a self-growing, self-organized neural gas network to calculate the most appropriate size of a codebook.

The resulting descriptor is formed by simultaneously employing each LF by two distinguished units. The first one is responsible for classifying the LF to a single class among the m ones calculated during training while assigning a participation value to it. The second unit describes the color of the LF's surrounding area using two fuzzy linking systems. This unit employs a 24-color palette to describe a color. The combination of the two units classifies the LF into at least one of the $m \times 24$ positions of the descriptor. Regarding the retrieval procedure the Term Frequency Inverse Document Frequency (TF-IDF) is used as the weighting scheme. The proposed recognition method has been chosen for the following reasons:

- It was tested in an object database and managed to present the best results among 15 descriptors
- The size of the codebook is automatically computed.
- To the best of our knowledge, this is the first method using color information in early fusion with visual words for object grapsing.
- Irrespectively of the database size, there is no need to consider weighted and/or similarity measure schemes.
- It exhibits good results in retrieving images from *long documents*, i.e. it can identify the presence of an object which matches to the query, even when it belongs to a cluttered image.
- It ensures high retrieval rates even in scaling changes and rotation variations.

One of the key issues in the design of such a grasping system is the database formation. Widely used databases were enriched with additional objects, one hundred instances of which were recorded in the database under controlled external conditions and different capturing viewpoints. The camera is placed at distance B from the object's centroid on the plane formed by the XY axes of the reference frame (placed at the center of the object) and is rotated with respect to the Z axis. Each object is captured every 36°, thus taking a total of 10 images. Next, the object is rotated clockwise by 36° on the Z axis and the camera repeats the same procedure as before. The overall routine concerning both the camera and the object is repeated, using a 36° step, until a complete rotation of the object around the Z axis is performed. It is therefore straightforward to note that, the smaller the rotation step, the better the results. The rotation step chosen is a trade-off between the database computational burden and the high retrieval rate scores.

During the system's operation, the object to be identified is captured and defined as the query object. The respective LFs are extracted and the vector describing the contents of the image is produced. The distance between this vector and the ones stored in the database is calculated. Since the database includes single objects, the query captured image, i.e. the one having the smaller distance from the database objects, is classified.

Please cite this article in press as: Kouskouridas, R., et al. What, Where and How? Introducing pose manifolds for industrial object manipulation. *Expert Systems with Applications* (2015), http://dx.doi.org/10.1016/j.eswa.2015.06.039

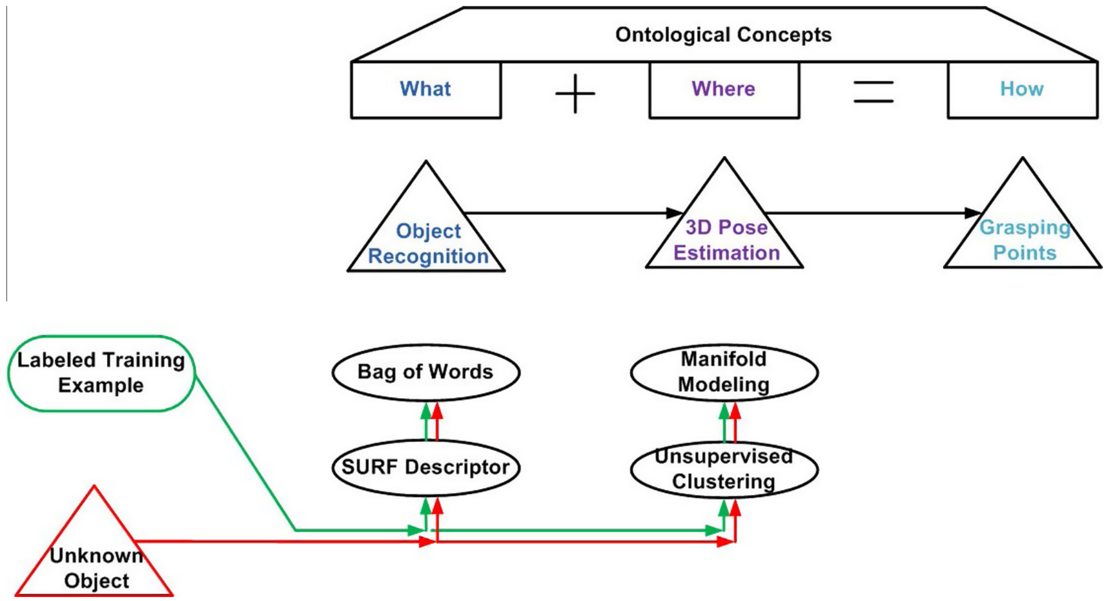

Fig. 2. The proposed methodology in a block diagram format.

3.2. 3D pose estimation – Where?

Our 6 DoF object pose estimation module can be apprehended as a generalization of the hypothesis presented in the previous section. The proposed method can be divided into two discriminative phases standing for the building of the part-based architecture and the manifold modeling one.

3.2.1. Part-based architecture

It is well understood that modeling objects as a collection of parts increases robustness to intraclass variation, pose change and even occlusion. The implicit shape model, introduced by Leibe, Leonardis, and Schiele (2004), learns, via unsupervised clustering, class-specific visual codebooks and spacial distributions for each entry. Codebook entries are then detected in the test image and used to cast probabilistic votes in the Hough space based on the learnt spatial distributions. Moreover, Gall, Yao, Razavi, Van Gool, and Lempitsky (2011) showed, with the class-specific Hough forest, how part-based modeling can be effectively combined with generalized Hough voting for object detection under the random forest framework.

We propose a novel method which compared to the aforementioned works requires less supervision and focuses on describing features representing both texture and geometrical attributes. Towards this end, SURF (Bay, Ess, Tuytelaars, & Gool, 2008) is used to abstract initial appearance-based characteristics, which are then processed by an homography-based RANSAC (Fischler & Bolles, 1981) to keep the most robust ones.

The geometrical attributes are aggregated by employing the $\mathcal{K}$-means algorithm over the locations of the texture-based features. First, we select b primary points from image I that contains an object o with pose p and mark them as $I(\boldsymbol{v}^b, o|p)^\rho$, $\boldsymbol{v} \in \mathbb{R}^2$. Furthermore, we keep only the locations $\boldsymbol{u} \in \mathbb{R}^2$ of the b extracted features. The latter form set $\mathcal{K}$ that is further processed by $\mathcal{K}$-means to calculate the respective clusters centroids that are from now on denoted as $\mathcal{S} = \langle\hat{\boldsymbol{\mu}}^{\mathcal{K}}, o|p\rangle$.

3.2.2. Manifold modeling – template matching

In theory manifold modeling and its further application of alignment, stands for a sophisticated approach to establish a similarity measure between two separate subspaces. As indicated by the benchmark work of Mei et al. (2011), objects when modeled as feature vectors of low dimensionality can be projected onto highly discriminative subspaces facilitating, thus, their accurate registration in the 3D environment. More recently, Pei, Huang, Shi, and Zha (2012) suggested how affine transformation can serve as a manifold-to-manifold distance measure to align the embedded motion patterns.

Compared to the state of the art of our manifold modeling architecture extract feature vectors of low dimensionality, i.e. 5 DoF's (location, scale, shape and orientation – similar to an ellipse in the Euclidean space). Let $\boldsymbol{r}$ represent the feature vector that spans our modeled manifold. Moreover, we assume that $\boldsymbol{r} = [\alpha, \beta, \gamma, \delta, \epsilon, \zeta]$, meaning that the members to be computed are equivalently represented by an ellipse $h(\boldsymbol{r}) = \alpha X^2 + \beta XY + \gamma Y^2 + \delta X + \epsilon Y + \zeta = 0$ in the Cartesian space. Here where (X, Y) corresponds to the collection of points of $h(\boldsymbol{r})$. To adequately fulfill the modeling process we propose a cost function minimization problem that is solved through PSO (Eberhart, Shi, & Kennedy, 2001):

$$H = \frac{1}{\mathcal{K}}\sum_{w=1}^{\mathcal{K}}\|\mathcal{S}-h\|^2 = \frac{1}{\mathcal{K}}\sum_{w=1}^{\mathcal{K}}\|\langle\hat{\boldsymbol{\mu}}^w, o|p\rangle - h(\boldsymbol{r})\|^2 + \lambda\sum_{j=1}^{5}(\boldsymbol{r}^j)^2 \qquad (1)$$

The last member of the cost function of Eq. (1) is a regularization factor experimentally set to $\lambda = 0.1$, which is added over α to ϵ (ζ is a bias).

Moreover, let f^τ represent the two foci of the estimated ellipse on the Cartesian space according to $f^\tau = \sqrt{\text{majoraxis}^2 - \text{minoraxis}^2}$. We model the pose manifold for object o with pose p as the L_2 distance between the extracted $\hat{\boldsymbol{\mu}}^{\mathcal{K}}$ clusters from the two foci of the ellipse:

$$\mathbf{x} = \|f-\mathcal{S}\|^2 = \sum_{\tau=1}^{2}\sum_{w=1}^{\mathcal{K}}\{f^\tau - \langle\hat{\boldsymbol{\mu}}^w, o|p\rangle\}^2 \qquad (2)$$

As a follow-up step we utilize a RBF-based regressor to find the correct mapping from a set of input variables $\mathbf{x} \in \mathcal{X}$ (pose space) to an output variable $\mathbf{y} = \mathbf{y}(\mathbf{x}; \theta) \in \mathcal{Y}$, where θ corresponds to the vector of the tunable parameters. The used datasets are CVL (Viksten, Forssén, Johansson, & Moe, 2009), COIL-100 (Nayar, Nene, & Murase, 1996), as well as a set of artificially rendered objects

Please cite this article in press as: Kouskouridas, R., et al. What, Where and How? Introducing pose manifolds for industrial object manipulation. *Expert Systems with Applications* (2015), http://dx.doi.org/10.1016/j.eswa.2015.06.039

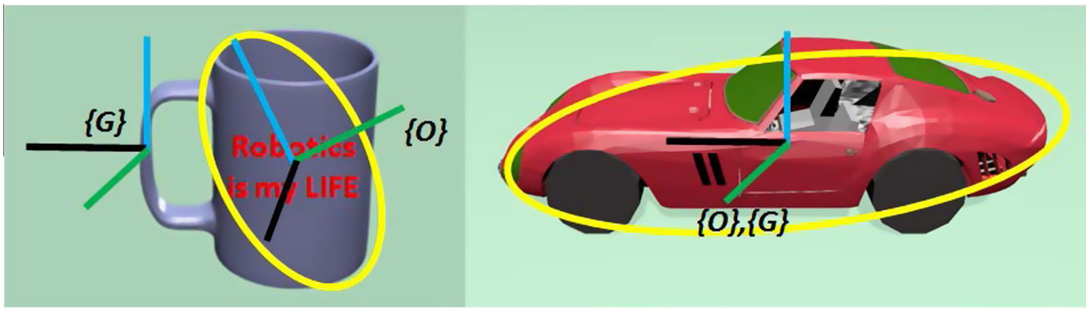

Fig. 3. Depending on the object class category, the grasping $\{G\}$ and the object frames $\{O\}$ might differ (e.g. a cup) or coincide (e.g. a car).

available on-line,[1] which are shot every 5°. In contrast to other related works, we do not utilize conventional dimensionality reduction schemes, e.g. PCA prone to the inevitable loss of information. The size of the training set is $[2 * \hat{\boldsymbol{\mu}}^{\mathcal{K}} \times 100,000]$, that is 1000 images/object. The number of the extracted clusters $\hat{\boldsymbol{\mu}}^{\mathcal{K}}$ was experimentally set to 8, which exhibited the lowest generalization error.

3.3. Grasping points – How?

Our grasping point calculation module takes into account information derived from both the recognition and 3D pose estimation frameworks. In most of the cases, an accurate estimation of the 3D pose of an object is sufficient for the ample accomplishment of manipulation tasks, since the robotic arm can be configured according to the provided 6 DoF measurements. However, in this paper we aim at enhancing our system by introducing grasping capabilities, so as a cup to be grasped by its handle but a toy-car about its center of mass. We believe that, this property increases the efficiency of the proposed system, making it appropriate for smart industrial applications. As Fig. 3 illustrates, the positions of the object frame $\{O\}$ and the grasping one $\{G\}$ are directly related to the class of the recognized object.

In more detail, the grasping points of cars and 4 legged animals are their center of mass, which are efficiently computed by finding the center of the respective fitted ellipse. On the other hand, cups and mugs imply different grasping points. Towards this end, let the transformation T_{CO}, describing the pose of the recognized object $\{O\}$ relatively to the camera frame $\{C\}$ to be denoted as:

$$T_{CO} = \begin{bmatrix} \mathbf{R}_{CO} & \mathbf{D}_{CO} \\ \mathbf{0}^T & 1 \end{bmatrix} \in SE(3)$$

where $\mathbf{R}_{CO}$ and $\mathbf{D}_{CO}$ represent the rotational and translational parameters (6 DoF) and are provided by the 3D pose estimation module presented in the previous Section 3.2. It is apparent that, an additional transformation T_{OG} is required to efficiently describe the spatial relationship between $\{O\}$ and $\{G\}$. Let $\mathbf{D}_{OG}$ and $\mathbf{R}_{OG}$ represent the translation and rotation matrix describing the orientation of the frame of the grasping point $\{G\}$ relatively to the object frame $\{O\}$. Then the transformation T_{OG} can be denoted as:

$$T_{OG} = \begin{bmatrix} \mathbf{R}_{OG} & \mathbf{D}_{OG} \\ \mathbf{0}^T & 1 \end{bmatrix} \in SE(3)$$

The recognition module provides specific grasping information that should be taken into account for the calculation of the grasping points. Apart from the identity of the sought object we hold data regarding its nature, i.e. the existence and the position of a handle, which is defined as the one that produces the best match along a series of comparisons for the same object class. Since the accurate location of the handle is known, we can calculate the $\mathbf{R}_{OG} \in SE(3)$ describing the orientation of the handle relative to the center of mass of the object. Finally, the transformation between the grasping point of the object and the camera is determined as:

$$T_{CG} = T_{CO} \cdot T_{OG} \tag{3}$$

3.4. Ontology-based grasping

The proposed architecture utilizes a knowledge-based information acquisition framework that consists of ontological concepts representing the three separate modules introduced so far. The employed ontology is structured as a graph, where each node represents an ontological concept and edges inter-relationships between them. An inter-relationship $\mathcal{E}$ between concepts D_i and D_j implies that there exists an inverse liaison $\mathcal{E}'$ between D_j and D_i.

A graphical representation of the proposed Ontology is depicted in Fig. 4, where the three general ontological concepts "*What*", "*Where*" and "*How*" are shown. Through the respective module of object recognition we determine the "*What*" ontological concept and its members (a testing object might either be a car, a cup or a 4-legged animal). Additionally, pose manifolds established via the respective pose estimation module, characterize the "*Where*" ontological concept that holds information regarding the 6 DoF geometrical configuration of the sought object. Finally, the attributes of the "*What*" and "*Where*" concepts are taken into account to generate the "*How*" hypothesis, which provides estimations about the grasping points of the respective objects. In this particular example, the testing target is firstly classified as a cup, which in turn, implies that its grasping point is derived through the ontology. Essentially, the proposed framework suggests that a novel object can be adequately manipulated after it has been initially recognized and afterwards assigned with a 6 DoF grasping point vector (see Fig. 5).

4. Experiments and discussion

The proposed framework was evaluated through a series of experiments to assess its performance in the particular task of grasping novel objects and its potential for industrial applications. Throughout these experiments we utilized "open loop" grasps that imply the uncontrolled movement of the robotic arm to the respective grasping point together with the closure of the fingers. Additionally, similar to Ulbrich et al. (2011), a candidate grasp is considered as successful in cases where it is a Force-Closure (FC) one, i.e. "*if and only if we can exert, though the set of contacts, arbitrary forces and moments on the object*" (Nguyen, 1986). Practically, Force-Closure grasps suggest that there exist equilibrium due to zero force and moments on the object, that in turn, can be translated as having the testing object grasped by the robotic gripper

[1] http://www.evermotion.org.

Please cite this article in press as: Kouskouridas, R., et al. What, Where and How? Introducing pose manifolds for industrial object manipulation. *Expert Systems with Applications* (2015), http://dx.doi.org/10.1016/j.eswa.2015.06.039

Fig. 4. (a) The "*What*", "*Where*" and "*How*" ontological concepts along with their interconnections; (b) A small portion of the constructed Ontology, where in this particular example, the Grasping_Point concept holds data regarding the 6 DoF pose of a cup in order to facilitate the efficient accomplishment of object manipulation tasks.

in a way that the likelihood of a fall to be minimized. We evaluated the performance of our approach on several novel objects belonging to the categories of cars, cups and 4-legged animals, respectively, whilst utilizing the SCORBOT-ER Vplus arm, a vertical articulated robot with 6 DoFs and a standard gripper. For the choice of the particular arm the main criterion was the existence of several industrial robotic systems with similar setup.

The efficacy of the proposed grasping point estimation module is directly related to the performance of the object recognition and pose estimation modules. Therefore, it is proper to state that in order to efficiently fulfill manipulation tasks, our method should cope with several computer vision challenges. Viewpoint changes and partial occlusions significantly affect the performance of the system. Contrary to humans, who are capable of simultaneously recognize and estimate the pose of a target in difficult conditions, robot vision applications fall short to achieve such robust responses. Towards this end, we have initially assessed the efficiency of the proposed method in cases where the testing object is either partially occluded or perceived by different perspectives. Experiments performed on the UkBekch database (Nister & Stewenius, 2006). Table 1 shows that our recognition module is capable of providing accurate estimations regardless of the viewing perspective and any partial occlusion, primarily due to the training with the BoF model. In these series of experiments the

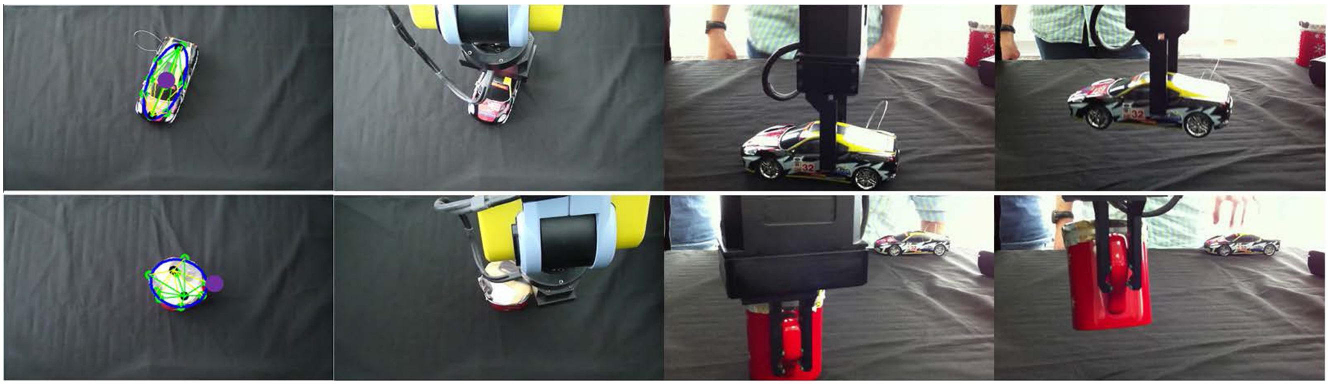

Fig. 5. The proposed method is capable of providing accurate grasping point estimations that enable the adequate manipulation of either a car or a cup. Here the fitted ellipse utilized through the manifold modeling process is shown with a blue line, while the purple dot is the grasping point computed.

Table 1
Precision levels of the proposed retrieval approach for objects observed under altering viewpoints or disturbed with partial occlusions.

Object	MAP	*P*@1	*P*@8
Cow	0.9665	1.0000	0.9330
Car	0.7998	0.9286	0.7210
Cup	0.7163	1.0000	0.6741

testing object is shot every 45°, rather than 36°, in order to assess the dynamic potential of the proposed framework. The 8 images of the object, without any partial occlusion, are considered to be the ground truth, which in turn lay in the range of [0–50]. Partial occlusions are generated with a black rectangle of random size being arbitrarily overlaid on the surface of the testing object. Additionally, query images with partial occlusions over the surface of the object were introduced to the proposed architecture to produce the results shown in Table 1. To evaluate the effectiveness of the proposed approach we used the precision-at-K ($P@K$, $K = 1$ for the first result and $K = 8$ for the first 8 positions) evaluation method as well as the Mean Average Precision (MAP) one. The trec files that include the detailed ranking lists of the experiments are available online. While other vision systems are highly affected by the common disturbances, our method is capable of providing accurate estimations about the grasping point of a novel object, thus enabling its adequate manipulation.

The comparative study and evaluation of different manipulation systems has always been of fundamental importance in the field of robotics research, while a sound benchmarking framework has not been realized yet. The variety of the hardware available, e.g. several robotic arms and hands, together with the application dependent software make a comparative evaluation almost impossible, still it can be coarsely distinguished into hardware-based and software-based studies (Michel, Bourquin, & Baillie, 2009). Regarding the first category, robotic arms and hands are evaluated through their recorded efficiency to facilitate complex manipulation tasks based on their dexterity and their DoFs. According to Michel et al. (2009) software-based, comparative testing is to be exhibited via simulation by means of assessing the efficacy of the respective algorithms in the particular task of object manipulation. In this paper, we adjust the technique of Michel et al. (2009) by performing a more qualitative rather than quantitative comparative evaluation of highly related contemporary systems. Towards this end, we introduce Table 2, where several comparison categories are apposed. The collated projects are analyzed based on the employed hardware, i.e. robotic hand/arm and vision sensor utilized and its architecture. Two major camera configurations for the scene perception by the robot are discerned, viz. the eye-in-hand and the eye-to-hand ones, which entail mounting the vision sensor(s) onto the robot's end-effector or installing the cameras separately from the robot, respectively. In the second case the camera should hold such a pose to provide the capability to observe the entire working space of the robot. Additionally, meticulous emphasis is given on whether the under consideration framework employs an object recognition module and on its generalization capabilities. Moreover, since systems that integrate simulation with other robotic platforms demonstrate higher force closure rates, in Table 2 we indicate whether the respective frameworks make use of either the GraspIt! (Miller & Allen, 2000) or the OpenGrasp (Ulbrich et al., 2011) environments. Finally, we

Table 2
Qualitative comparison of related frameworks.

Framework	Robotic hand/arm utilized	Object recognition/generalization capabilities	Simulation with other platforms	Grasping point estimation/generalization capabilities	Vision sensor utilized	Camera configuration
Hsiao et al. (2010)	PR2 personal robot	No	GraspIt!	Supervised/limited	Stereo cameras	Eye-to-hand
Curtis and Xiao (2008)	Barrett hand/PUMA arm	No	GraspIt!	Unsupervised/high	n/a	n/a
Madry et al. (2012)	Barrett hand/PUMA arm	Yes/high	No	Supervised/limited	Stereo cameras	Eye-to-hand
Chiu et al. (2010)	Barrett hand/Barrett arm	Yes/limited	No	Supervised/limited	Monocular	Eye-to-hand
Proposed method	**SCORBOT-ER Vplus arm**	**Yes/high**	**GraspIt! OpenGrasp**	**Unsupervised/high**	**Monocular**	**Both**
Boularias et al. (2011)	Barrett hand/Mitsubishi PA-10 arm	No	GraspIt!	Supervised/high	Monocular time-of-flight	Eye-to-hand
Richtsfeld and Vincze (2011)	OttoBock hand/AMTEC arm	Yes/limited	No	Supervised/limited	Monocular	Eye-to-hand
Saxena et al. (2008)	STAIR I/STAIR II	No	No	Supervised/high	Stereo cameras	Eye-to-hand
Huang et al. (2012)	PUMA 500	Yes/limited	No	Supervised/limited	Monocular	Eye-in-hand

qualitatively compare the grasping point estimation modules of the respective framework in a way to draw meaningful conclusions regarding their ability to be adopted by other techniques or to generalize to other objects.

A grasp selection method that makes use of 3D sensor data to appoint a ranked set of potential grasps for an object placed on a workbench at a predefined location is proposed in Hsiao, Chitta, Ciocarlie, and Jones (2010). Stereo cameras installed on the PR2 robot are used and the simulation results are obtained through the GraspIt! environment. Additionally, although the recorded rates for "open loop" grasps are impressive, the limited set of testing objects along with the restricted experimental justification suggest that this technique is very unlikely to reproduce similar results. In Curtis and Xiao (2008), a work that shares common spirit with this paper is presented, in the sense that it incorporates a knowledge transfer module to facilitate the efficient manipulation of novel objects. However, their method does not implement an object recognition framework since it emphasizes only in classifying the testing objects into categories that correspond to known geometrical shapes. Grasping points are learnt through an iterative and interactive process and characterize the entire shape class, thus offering large generalization capacities. Despite its sophisticated architecture and simulation via GraspIt!, this method fails to enable the adequate manipulation of a cup or a mug through its handle, whilst providing mere speculations regarding the vision sensors used. The works in Madry, Song, and Kragic (2012) and Chiu, Liu, Kaelbling, and Lozano-Pérez (2010) employ an object recognition framework to empower their respective grasping point estimation method, while regarding the pattern identification module, the method of Madry et al. (2012) presented higher generalization capacities. Additionally, both systems utilize eye-to-hand camera architectures, whilst their grasping point selection frameworks fail to generalize to novel objects belonging to different classes.

In order to fully assess the performance of the proposed framework in the particular task of manipulating novel objects, we run several tests that include simulation with other robotic hands and grippers, rather the SCORBOT-ER Vplus arm. Through these tests, we artificially generated 3D object models belonging to the classes of a car, cup or 4-legged animals and provided to the virtual controller the respective vision algorithm that acquired images through the Frame Grabber subroutine (in the GraspIt! environment). Our method enabled the adequate accomplishment of object manipulation tasks through its recognition module that facilitates the transfer of grasping point-based knowledge. The utilized ontology architecture provides high generalization capacities to the grasping point selection module and minimizes the complexity of the framework. Evaluation with real objects in realistic scenarios as those depicted in the accompanying video, provide evidence of high force closure grasping rates and low generalization error. Cars and 4-legged animals are correctly grasped at their center of mass while cups are manipulated from their handle. Throughout the literature, the only method that offers similar capacities is reported in Saxena, Driemeyer, and Ng (2008), where the STAIR I and II robots are utilized to perform demanding pick and place tasks. However, an object recognition subroutine is not suggested in Saxena et al. (2008), albeit their supervised grasping point estimation process offers high generalization capacities.

The most trivial solution to the unknown object manipulation problem is presented in Richtsfeld and Vincze (2011), where images of known models are taken into account to form a 3D representation of the testing scene. This enables the accurate estimation of the respective grasping points of the objects, limiting, however, the object operation range of the method to trained models only. The most recently proposed systems are those presented in Boularias, Kroemer, and Peters (2011) and Huang, Walker, and Birchfield (2012). The advent of technology enabled the use of Time-of-flight cameras in Boularias et al. (2011) for the acquisition of accurate 3D data, which in turn, were fused with shape information retrieved from the Princeton Shape Benchmark. Similar to (Curtis & Xiao, 2008), sophisticated shape descriptions provide evidence of low generalization error however, they fail to appoint the grasping point of a cup as its handle. It is apparent that, all the aforementioned frameworks rely on highly sophisticated modules that are independent of the utilized hardware. The proposed hardware-independent solution, through the class-dependent pose manifolds and its novel object recognition module, provided invariance to large displacements and partial occlusions, whilst easing the grasping of unregistered objects.

5. Conclusion and future work

In this paper an integrated framework for industrial object manipulation was presented. Our method suggests a novel solution to the automatic manipulation of objects for industrial purposes in terms of providing a low complexity architecture capable of generalizing to unknown objects without requiring additional learning of new objects. The system exhibits real-time performance and it can be easily adopted by any robotic platform regardless of its components, e.g. gripper, joints, etc. Moreover, the presented framework integrates ontological models into a unified context for the particular task of the autonomous object grasping. It ranges from a perceptual object recognition module up to a semantic based categorization of object affordances. We believe that one of the major challenges in industrial-centric object grasping is trying to answer the three questions of *What*, *Where*, and *How*, in a cohesive way, without leaving out critical semantic constraints that are affecting the whole context of the object manipulation tasks. The proposed system, addresses the recognition problem via a BoF classification scheme from a shape based perspective. A 6 DoF pose estimation technique incorporates a robust bunch-based architecture along with a manifold modeling procedure. The grasping points are then identified through an ontology-based knowledge acquisition, where the recognized objects inherit their affordances from the respective classes. In such a way, an ontologization concept is realized focusing on the liaison between computer vision algorithms and the corresponding motor commands to accomplish grasping of an unknown object. Unlike other contemporary solutions, which either crave labor-intensive on-line learning or construct high dimensional input vectors, the proposed method requires minimum supervision and low dimensionality training data, thus minimizing the complexity of the system and making it appropriate for industrial applications. We believe that our vision-based solution for the particular problem addresses all the challenging issues and offers high adaptability and large generalization capabilities in a minimum cost.

With an outlook to the future work, we consider replacing the Bag-of-Visual-Words model with the Vector of Locally Aggregated Descriptors (VLAD) (Jegou, Douze, Schmid, & Pérez, 2010) model. Given a codebook, instead of creating a vector of frequencies, the VLAD model produces a vector of differences, as distances, between a feature and the clusters center. This approach significantly reduces the number of the codebook clusters to tiny sizes while maintaining robust performances. Moreover, in order to highlight the effectiveness of our approach it is important to perform experiments using larger set of objects. Furthermore, we can also repeat the experiments employing additional objects in the training set as distractors, to assess the large-scale recognition performance. Finally, following the recent advances in deep learning we also consider designing a deep Convolutional Neural Network or a deep network of sparse auto-encoders to automatically learn highly discriminative features for object recognition and pose estimation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.eswa.2015.06.039.

References

Aly, M., Welinder, P., Munich, M. E., & Perona, P. (2009). Automatic discovery of image families: Global vs. local features. In *Proceedings of the international conference on image processing, ICIP 2009, 7–10 November 2009, Cairo, Egypt* (pp. 777–780).

Arampatzis, A., Zagoris, K., & Chatzichristofis, S. A. (2013). Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management, 49*, 274–285. http://dx.doi.org/10.1016/j.ipm.2012.03.005.

Bannat, A., Bautze, T., Beetz, M., Blume, J., Diepold, K., Ertelt, C., et al. (2011). Artificial cognition in production systems. *IEEE Transactions on Automation Science and Engineering, 8*, 148–174.

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. J. V. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding, 110*, 346–359.

Bicchi, A. (2000). Hands for dexterous manipulation and robust grasping: A difficult road toward simplicity. *IEEE Transactions on Robotics and Automation, 16*, 652–662.

Bonde, U., Badrinarayanan, V., & Cipolla, R. (2014). Robust instance recognition in presence of occlusion and clutter. In *ECCV*.

Boularias, A., Kroemer, O., & Peters, J. (2011). Learning robot grasping from 3-d images with Markov random fields. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (pp. 1548–1553).

Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., & Rother, C. (2014). Learning 6d object pose estimation using 3d object coordinates. In *ECCV*.

Buchholz, D., Futterlieb, M., Winkelbach, S., & Wahl, F. M. (2013). Efficient bin-picking and grasp planning based on depth data. In *2013 IEEE international conference on robotics and automation (ICRA)* (pp. 3245–3250). IEEE.

Buchholz, D., Kubus, D., Weidauer, I., Scholz, A., & Wahl, F. M. (2014). Combining visual and inertial features for efficient grasping and bin-picking. In *2014 IEEE international conference on robotics and automation (ICRA)* (pp. 875–882). IEEE.

Cao, X., Ning, B., Yan, P., & Li, X. (2012). Selecting key poses on manifold for pairwise action recognition. *IEEE Transactions on Industrial Informatics, 8*, 168–177.

Chatzichristofis, S., Iakovidou, C., Boutalis, Y., & Marques, O. (2013). Co.vi.wo.: Color visual words based on non-predefined size codebooks. *IEEE Transactions on Cybernetics, 43*, 192–205. http://dx.doi.org/10.1109/TSMCB.2012.2203300.

Chatzistavros, E., Chatzichristofis, S. A., Zagoris, K., & Stamatelos, G. (2015). Content-based image retrieval over ieee 802.11b noisy wireless networks. *International Journal of Communication Systems*, 1432–1449. http://dx.doi.org/10.1002/dac.2724.

Chen, S., Li, Y., & Kwok, N. M. (2011). Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research, 30*, 1343–1377.

Chiu, H., Liu, H., Kaelbling, L., & Lozano-Pérez, T. (2010). Class-specific grasping of 3d objects from a single 2d image. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (pp. 579–585).

Curtis, N., & Xiao, J. (2008). Efficient and effective grasping of novel objects through learning and adapting a knowledge base. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (pp. 2252–2257).

Da Xu, L., Wang, C., Bi, Z., & Yu, J. (2012). Autoassem: An automated assembly planning system for complex products. *IEEE Transactions on Industrial Informatics, 8*, 669–678.

Drost, B., Ulrich, M., Navab, N., & Ilic, S. (2010). Model globally, match locally: efficient and robust 3d object recognition. In *CVPR*.

Eberhart, R., Shi, Y., & Kennedy, J. (2001). Swarm intelligence. In *The Morgan Kaufmann series in evolutionary computation*.

Ferrari, V., Tuytelaars, T., & Van Gool, L. (2006). Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision, 67*, 159–188.

Fischler, M., & Bolles, R. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*, 381–395.

Gall, J., Yao, A., Razavi, N., Van Gool, L., & Lempitsky, V. (2011). Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*, 2188–2202.

Gruber, T. et al. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199–220.

Hinterstoisser, S., Benhimane, S., & Navab, N. (2007). N3m: Natural 3d markers for real-time object detection and pose estimation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1–7).

Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., & Lepetit, V. (2011). Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *ICCV*.

Hsiao, K., Chitta, S., Ciocarlie, M., & Jones, E. (2010). Contact-reactive grasping of objects with partial shape information. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (Vol. 10, p. 2010).

Huang, X., Walker, I., & Birchfield, S. (2012). Occlusion-aware reconstruction and manipulation of 3d articulated objects. In *Proceedings of the IEEE international conference on robotics and automation* (pp. 1365–1371).

Hu, B., Dasmahapatra, S., Lewis, P., & Shadbolt, N. (2003). Ontology-based medical image annotation with description logics. In *Proceedings of the IEEE international conference on tools with artificial intelligence* (pp. 77–82).

Iakovidou, C., Anagnostopoulos, N., Kapoutsis, A. C., Boutalis, Y. S., & Chatzichristofis, S. A. (2014). Searching images with MPEG-7 (& mpeg-7-like) powered localized descriptors: The SIMPLE answer to effective content based image retrieval. In *2014 12th international workshop on content-based multimedia indexing (CBMI), Klagenfurt, Austria, June 18–20, 2014* (pp. 1–6). http://dx.doi.org/10.1109/CBMI.2014.6849821.

Jegou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *The twenty-third IEEE conference on computer vision and pattern recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010* (pp. 3304–3311).

Kostavelis, I., & Gasteratos, A. (2013). Learning spatially semantic representations for cognitive robot navigation. *Robotics and Autonomous Systems, 61*, 1460–1475.

Kouskouridas, R., Charalampous, K., & Gasteratos, A. (2014). Sparse pose manifolds. *Autonomous Robots, 37*, 191–207.

Kouskouridas, R., Amanatiadis, A., & Gasteratos, A. (2011). Guiding a robotic gripper by visual feedback for object manipulation tasks. In *Proceedings of the IEEE international conference on mechatronics* (pp. 433–438).

Kouskouridas, R., Retzepi, T., Charalampoglou, E., & Gasteratos, A. (2012). Ontology-based 3d pose estimation for autonomous object manipulation. In *2012 IEEE international conference on imaging systems and techniques (IST)* (pp. 476–481). IEEE.

Kragic, D., & Christensen, H. (2003). Robust visual serving. *International Journal of Robotics Research, 22*, 923–939.

Kuo, H.-Y., Su, H.-R., Lai, S.-H., & Wu, C.-C. (2014). 3d object detection and pose estimation from depth image for robotic bin picking. In *2014 IEEE international conference on automation science and engineering (CASE)* (pp. 1264–1269). IEEE.

Leibe, B., Leonardis, A., & Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *Workshop, ECCV*, (pp. 17–32).

Lim, J. J., Khosla, A., & Torralba, A. (2014). Fpm: Fine pose parts-based model with 3d cad models. In *ECCV*.

Liu, H. (2011). Exploring human hand capabilities into embedded multifingered object manipulation. *IEEE Transactions on Industrial Informatics, 7*, 389–398.

Liu, M.-Y., Tuzel, O., Veeraraghavan, A., Taguchi, Y., Marks, T. K., & Chellappa, R. (2012). Fast object localization and pose estimation in heavy clutter for robotic bin picking. *The International Journal of Robotics Research, 31*, 951–973.

Ma, J., Chung, T., & Burdick, J. (2011). A probabilistic framework for object search with 6-dof pose estimation. *International Journal of Robotics Research, 30*, 1209–1228.

Madry, M., Song, D., & Kragic, D. (2012). From object categories to grasp transfer using probabilistic reasoning. In *Proceedings of the IEEE international conference on robotics and automation* (pp. 1716–1723).

Manjunath, B. S., Ohm, J. R., Vasudevan, V. V., & Yamada, A. (2001). Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology, 11*, 703–715.

McGuire, P., Fritsch, J., Steil, J., Rothling, F., Fink, G., Wachsmuth, S., Sagerer, G., & Ritter, H. (2002). Multi-modal human–machine communication for instructing robot grasping tasks. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (pp. 1082–1088).

Mei, L., Liu, J., Hero, A., & Savarese, S. (2011). Robust object pose estimation via statistical manifold modeling. In *Proceedings of the IEEE international conference on computer vision* (pp. 967–974).

Mezaris, V., Kompatsiaris, I., & Strintzis, M. (2003). An ontology approach to object-based image retrieval. In *Proceedings of the IEEE international conference on image processing* (Vol. 2, pp. 511–514).

Michel, O., Bourquin, Y., & Baillie, J. (2009). Robotstadium: Online humanoid robot soccer simulation competition. In *RoboCup 2008: Robot soccer world cup XII*, (pp. 580–590).

Miller, A., & Allen, P. (2000). Graspit!: A versatile simulator for grasp analysis. In *Proceedings of the of the ASME dynamic systems and control division*.

Modayil, J., & Kuipers, B. (2007). Autonomous development of a grounded object ontology by a learning robot. *Proceedings of the national conference on artificial intelligence* (Vol. 22, pp. 1095). Menlo Park, CA; Cambridge, MA: AAAI Press. MIT Press, London; 1999.

Nayar, S., Nene, S., & Murase, H. (1996). Columbia Object Image Library (COIL 100). Technical Report Tech. Report No. CUCS-006-96. Department of Comp. Science, Columbia University.

Nguyen, V. (1986). *The synthesis of stable force-closure grasps* (Master's thesis MIT). Dept. of Electrical Engineering and Computer Science.

Nieuwenhuisen, M., Droeschel, D., Holz, D., Stuckler, J., Berner, A., Li, J., et al. (2013). Mobile bin picking with an anthropomorphic service robot. In *2013 IEEE international conference on robotics and automation (ICRA)* (pp. 2327–2334). IEEE.

Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 2, pp. 2161–2168).

Papadopoulos, D. P., Kalogeiton, V. S., Chatzichristofis, S. A., & Papamarkos, N. (2013). Automatic summarization and annotation of videos with lack of metadata information. *Expert Systems with Applications, 40*, 5765–5778. http://dx.doi.org/10.1016/j.eswa.2013.02.016.

Pei, Y., Huang, F., Shi, F., & Zha, H. (2012). Unsupervised image matching based on manifold alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 1658–1664.

Popovic, M., Kraft, D., Bodenhagen, L., Baseski, E., Pugeault, N., Kragic, D., et al. (2010). A strategy for grasping unknown objects based on co-planarity and colour information. *Robotics and Autonomous Systems, 58*, 551–565.

Ren, R., Collomosse, J., & Jose, J. (2011). A bovw based query generative model. In *Proceedings of the international conference on advances in multimedia modeling* (pp. 118–128).

Richtsfeld, M., & Vincze, M. (2011). Robotic grasping of unknown objects. Robot Arms. URL <http://www.intechopen.com/books/export/citation/BibTex/robot-arms/robotic-grasping-of-unknown-objects1>.

Rios-Cabrera, R., & Tuytelaars, T. (2013). Discriminatively trained templates for 3d object detection: A real time scalable approach. In *2013 IEEE international conference on computer vision (ICCV)* (pp. 2048–2055). IEEE.

Sansoni, G., Bellandi, P., Leoni, F., & Docchio, F. (2014). Optoranger: A 3d pattern matching method for bin picking applications. *Optics and Lasers in Engineering, 54*, 222–231.

Saxena, A., Driemeyer, J., & Ng, A. (2008). Robotic grasping of novel objects using vision. *International Journal of Robotics Research, 27*, 157–173.

Song, S., & Xiao, J. (2014). Sliding shapes for 3d object detection in depth images. In *ECCV*.

Steil, J., Röthling, F., Haschke, R., & Ritter, H. (2004). Situated robot learning for multi-modal instruction and imitation of grasping. *Robotics and Autonomous Systems, 47*, 129–141.

Tejani, A., Tang, D., Kouskouridas, R., & Kim, T.-K. (2014). Latent-class hough forests for 3d object detection and pose estimation. In *Computer vision – ECCV 2014* (pp. 462–477). Springer.

Ulbrich, S., Kappler, D., Asfour, T., Vahrenkamp, N., Bierbaum, A., Przybylski, M., & Dillmann, R. (2011). The opengrasp benchmarking suite: An environment for the comparative analysis of grasping and dexterous manipulation. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems* (pp. 1761–1767).

Varadarajan, K. M., & Vincze, M. (2012). Afrob: The affordance network ontology for robots. In *2012 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1343–1350). IEEE.

Viksten, F., Forssén, P., Johansson, B., & Moe, A. (2009). Comparison of local image descriptors for full 6 degree-of-freedom pose estimation. In *Proceedings of the IEEE international conference on robotics and automation* (pp. 2779–2786).

Vorobieva, H., Soury, M., Hède, P., Leroux, C., & Morignot, P. (2010). Object recognition and ontology for manipulation with an assistant robot. *Aging Friendly Technology for Health and Independence*, 178–185.

Wang, L., Ren, L., Mills, J. K., & Cleghorn, W. L. (2010). Automated 3-d micrograsping tasks performed by vision-based control. *IEEE Transactions on Automation Science and Engineering, 7*, 417–426.

Wang, G., Tao, L., Di, H., Ye, X., & Shi, Y. (2012). A scalable distributed architecture for intelligent vision system. *IEEE Transactions on Industrial Informatics, 8*, 91–99.

Wohlhart, P., & Lepetit, V. (2015). Learning descriptors for object recognition and 3d pose estimation. In *CVPR*.

Ying, L., Fu, J., & Pollard, N. (2007). Data-driven grasp synthesis using shape matching and task-based pruning. *IEEE Transactions on Visualization and Computer Graphics, 13*, 732–747.