Auton Robot (2014) 37:191–207
DOI 10.1007/s10514-014-9388-x

Sparse pose manifolds

Rigas Kouskouridas · Kostantinos Charalampous · Antonios Gasteratos

Received: 1 March 2013 / Accepted: 25 February 2014 / Published online: 25 March 2014

Abstract The efficient manipulation of randomly placed objects relies on the accurate estimation of their 6 DoF geometrical configuration. In this paper we tackle this issue by following the intuitive idea that different objects, viewed from the same perspective, should share identical poses and, moreover, these should be efficiently projected onto a well-defined and highly distinguishable subspace. This hypothesis is formulated here by the introduction of pose manifolds relying on a bunch-based structure that incorporates unsupervised clustering of the abstracted visual cues and encapsulates appearance and geometrical properties of the objects. The resulting pose manifolds represent the displacements among any of the extracted bunch points and the two foci of an ellipse fitted over the members of the bunch-based structure. We post-process the established pose manifolds via l_1 norm minimization so as to build sparse and highly representative input vectors that are characterized by large discrimination capabilities. While other approaches for robot grasping build high dimensional input vectors, thus increasing the complexity of the system, in contrast, our method establishes highly distinguishable manifolds of low dimensionality. This paper represents the first integrated research endeavor in formulating sparse pose manifolds, with experimental results providing evidence of low generalization error, justifying thus our theoretical claims.

Keywords Object manipulation · Sparse representation · Manifold modeling · Neural networks

R. Kouskouridas (✉)
Computer Vision & Learning Lab, Department of Electrical & Electronic Engineering, Imperial College London, South Kensington, London SW7 2AZ, UK
e-mail: r.kouskouridas@imperial.ac.uk

K. Charalampous · A. Gasteratos
Laboratory of Robotics & Automation, Department of Production & Management Engineering, Democritus University of Thrace, Vasilissis Sophias 12, GR-671 00 Xanthi, Greece
e-mail: kchara@pme.duth.gr

A. Gasteratos
e-mail: agaster@pme.duth.gr

1 Introduction

The performance of any robotic devices capable of performing manipulation tasks is directly related to its ability not only to recognize objects, but to also provide accurate estimations about their 3D geometrical configuration. Object manipulation is a very complex procedure involving the accurate estimation of the 6 DoF pose of the testing target, the efficient avoidance of possible obstacles in the arm's path and the design of a grasping strategy for hand(finger)-object alignment (Saxena et al. 2011). Most human operations are related to object manipulation, such as drinking, eating, using a tool, door opening, driving, etc. Driven by its fundamental importance and the ability to be adopted by an abundance of diverse applications, several research endeavors were directed towards object manipulation (Rasolzadeh et al. 2010; Mason et al. 2011; Lippiello et al. 2011; Oikonomidis et al. 2011; Saxena et al. 2008, 2006) or object searching (Shubina and Tsotsos 2010; Andreopoulos and Tsotsos 2009). In recent years, scholarly activity has focused on solving object recognition and 3D pose estimation through methods involving the construction of large databases of images of objects to be compared and matched with similar ones during the testing phase (Detry and Piater 2011; Hsiao et al. 2010; Ferrari et al. 2006; Kouskouridas et al. 2012; Popovic et al. 2010). However, such solutions fail to attract stakeholders' interest since, in most of the cases, a robot's working environment could well contain unknown objects positioned

randomly. Towards this end, emphasis is given on designing and implementing novel robust image understanding techniques that can solve the problem of grasping unknown objects under minimum supervision. However, there is no ground solution to the manipulation problem combining large generalization capacities together with modest computational complexity, despite the substantial endeavors and certain achievements so far.

This paper proposes a generalized solution to the 6 DoF object pose estimation problem, with view to solve manipulation tasks, assuming an obstacle-free path and a straightforward grasping strategy. We believe that, under these assumptions, our approach represents a more general solution to object grasping that could be easily adopted by any robotic architecture. The proposed *Sparse Pose Manifolds* (SPM) method aims at solving the problem of unknown object manipulation by unifying 3D pose manifolds and grasping points into a cohesive framework for robot grasping, based on the following intuitive ideas: (a) an object observed under shifting viewpoints holds properties that can be sufficiently modeled and projected onto a well-defined subspace, highly distinguishable among its neighbors, and (b) even totally different objects captured under similar perspectives should share identical 3D pose attributes, leading to similar measurements that can indeed establish highly discriminative manifolds, capable of assigning accurate 3D pose tags to different object models. The proposed pose manifolds, formulated via a sophisticated architecture, stand for a low dimensional data representation that enables direct modification of the configuration of the robot's wrist according to the estimated grasping points. Unlike contemporary systems requiring extensive supervision and large repositories of images of objects, our focus is on providing a ground solution, with large generalization capacities based on unsupervised learning. Towards this end, we employ manifold modeling procedures that emphasize on processing available visual data in such a way so that their projection onto the corresponding subspaces is sparse, compact and highly representative. Figure 1 illustrates the basic ideas upon which the proposed framework is built. Contrary to the existing solutions for object manipulation, requiring *a priori* knowledge of the geometry of the object, our method builds distinguishable pose manifolds that categorize identical poses of different objects into the same classes.

Initially, given an image of an object with a specific pose, we derive a feature vector through the proposed ellipse-fitting module. This procedure iterates over the entire dataset comprising of object images belonging to 6 different classes. We then employ l_1 norm minimization in order to derive sparse representation vectors and, henceforth, to enhance the generalization capabilities of our method. As a follow-up step, the derived manifolds are mapped, through supervised learning, to a well-defined and highly representative subspace. The ultimate goal of the proposed module is to efficiently estimate the 6 DoF pose of an unknown testing object with view to facilitate grasping operations. After the efficient estimation of the pose of the object, through the manifold modeling procedure, its geometrical attributes (pose relative to the robots frame) are given as input to the robots controller to perform a grasp.

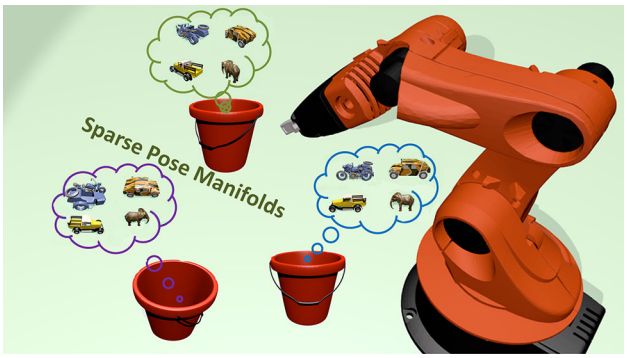


Fig. 1 The key idea underlying the proposed framework is the establishment of highly representative manifolds that lie on subspaces of low dimensionality. Different objects shot under the same aspects share identical 6DoF estimations that are categorized into the corresponding classes to form the pose manifolds. In this figure, these classes are represented by buckets of a specific pose, whilst the ultimate goal of the proposed visual grasping method is to efficiently guide the gripper to grasp an unknown object

It is our belief that this work of ours makes a serious contribution to the present state of the art, in the following ways:

- The proposed *Sparse Pose Manifolds* method lays its foundations on a novel module for pose manifold modeling that establishes compact and highly representative subspaces, whilst experimental verification provide evidence for the existence of such manifolds, sufficient to facilitate unknown object grasping.
- Pose manifolds depend on a novel bunch-based architecture (here introduced for the first time) which, unlike previous works (Savarese and Fei-Fei 2007; Hinterstoisser et al. 2007), is able to bypass the part selection process using unsupervised clustering, whereas by extracting local patches, encapsulates both appearance and geometrical attributes of the objects.
- Extending earlier work reported in Mei et al. (2009, 2011) where 3D pose estimations are limited to cars, we utilize numerous databases of real objects available that are further expanded by new artificially generated ones. In addition, compared to previous work our method offers higher generalization capacities mainly due to the efficient learning based on a large *a priori* training set containing numerous examples of real and artificial data.

- The established manifolds exhibit low dimensionality, thus avoiding the usage of conventional dimensionality reduction schemes widely employed in several computer vision applications. Besides, through the minimization of the l_1 norm we build sparse and compact manifolds that are highly representative, tolerant to common imaging disturbances (e.g. noise) and superior in terms of discrimination capabilities.
- Comparative evaluation of our work against other related works demonstrates its superiority in 3D pose estimation and object manipulation tasks. Quantitative and qualitative experimental results provide evidence of high 3D pose recovery rates with low generalization error, whist real time execution boost the potential of our work.

This paper is organized as follows. In the following Subsection 1.1, we provide an overview of the related work in the fields of eye-in(to)-hand target manipulation, sparse feature representation and 3D pose estimation. In Sect. 2 we formulate the proposed *Sparse Pose Manifolds* (SPM) method. Experimental results demonstrating the quality of our work are presented in Sect. 3, while some final notes and conclusions are drawn in Sect. 4.

1.1 Related work

An essential phase in any object classification and pattern recognition task is the feature extraction procedure. Most classifiers attempt to compute a separating hyper-plane between the corresponding classes, which leads to a variety of published methodologies (Bishop 2006). Considering the simple two-class problem, in which a linear discriminant function should be computed, the decision hyperplane is $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$, where $\mathbf{w}$ is the weight vector, w_0 is the bias term and $\mathbf{x}$ is a sample vector. However, the representation of an object in a low dimensional space aiming to provide a more discriminative information is the main objective of any feature extraction technique, as the initial step towards a minimum error classification and generalization capacity. Feature extraction algorithms are divided into three categories, namely unsupervised, supervised and semi-supervised ones, depending on the provision of the labels of the corresponding samples. The majority of those methods seek projection vectors $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_l]$ that minimize a particular cost function, following a certain criterion, ultimately supplying an optimized low dimensional representation of a dataset $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_n}]$. Regarding the unsupervised methodologies, the Principal Component Analysis (PCA) (Jolliffe 1986) , in which the optimization function seeks a sub-space, in order to maximize the variance of the data, is considered the most popular. Linear Discriminant Analysis (Duda et al. 2001) however, another well-known supervised technique, attempts to maximize the ratio of the between-class to within-class scatter matrices. Furthermore, the non-linear extension induced via support vector theory (Schölkopf and Smola 2002), known as "kernel trick', lead to many kernelized versions of linear approaches, such as Kernel-PCA (Schölkopf et al. 1997) and Generalized Discriminant Analysis (GDA) (Baudat and Anouar 2000). A newly developed framework is the Graph Embedding (GE) (Yan et al. 2007), in which all data samples are considered to form a graph $G = \{X, Q\}$, where Q is a similarity matrix incorporating either geometrical or statistical relations that prevail the data set. The latter yields a low dimensional manifold encapsulating the relations in the initial high-dimensional space. Many known techniques can be clarified under GE, and depending on the definition of the corresponding similarity measure, the resulting manifold may be identical to that produced by another algorithm, such as the PCA. In other words, GE encapsulates the majority of feature extraction methods. Sparse representation has been introduced in signal processing and pattern recognition via l_1 norm minimization, which can be solved using standard linear programming (Chen et al. 2001). The inherent properties that the l_1 norm makes available, including discrimination capability and noise tolerance, have drawn the attention of scholars in computer vision and pattern recognition fields (Qiao et al. 2010). The l_1 norm minimization has already been exploited in graph techniques (Cheng et al. 2010), dimensionality reduction (Zou et al. 2004; Pang et al. 2010), classification (Wright et al. 2009), and action recognition (Guha and Ward 2012; Wang et al. 2013; Castrodad and Sapiro 2012) leading to optimum results.

Based on the literature, in the field of robotics research two major camera configurations for vision-based scene understanding are discerned. The eye-in-hand and the eye-to-hand architectures entail mounting the vision sensor(s) onto the robot's end-effector or installing the cameras separate from the robot, possessing however such a pose with the capability to observe its entire working space. Such configurations are utilized in several applications providing solutions to the problems of 3D object modeling (Torabi and Gupta 2012; Krainin et al. 2011), visual servoing (Chan et al. 2011; Lippiello et al. 2007) and object manipulation (Agrawal et al. 2010; Kragic et al. 2005; Ben Amor et al. 2012; Hebert et al. 2012; Bohg and Kragic 2009). Srinivasa et al. (2010) proposed a 6 DoF pose estimation technique that shares common ideas with the work of Choi et al. (2008), while incorporating the process of matching images of objects of the test scene with 3D object models of a database constructed offline. Two eye-to-hand cameras allow stereo-based depth analysis which, when combined with data transmitted from the eye-in-hand sensor, is accurate enough for 3D pose measurement of the testing patterns. Besides, feature extraction is accomplished via the Scale Invariant Feature Transform (SIFT) whilst, for the fine adjustment of pose parameters,

the problem is formulated as an optimization one and solved by the Levenberg-Marquardt algorithm.

The term pose / grasping manifolds was first coined in Tsoli and Jenkins (2008) with a view to propose a manifold construction scheme based on data representing the motion of a human hand during an object manipulation task. However, this approach corresponds to a rather trivial method that primarily builds high dimensional manifolds, whilst employing conventional dimensionality reduction frameworks for subspace embedding. Furthermore, other drawbacks of the work presented in Tsoli and Jenkins (2008) are thought to be: (a) the configuration of the arm, obtained through a pose simulation that needs to be provided by the user; and (b) lack of solution to the grasping point calculation problem. In contrast, in this paper the term pose manifold is adopted with the view to indicate that there is a direct relation between pose-related data and the desired arrangement of the robot's end-effector.

Mei et al. (2009) have shown that given a highly representative dataset and an adequate manifold modeling subroutine, the performance of 3D object pose estimation is appropriately bootstrapped to the compactness of the established manifolds. In addition, they proved that part-based architectures, unlike holistic ones, are capable of handling more efficiently usual imaging disturbances such as partial occlusions and background clutter. The key idea underlying the part-based methods is the pursuit of points or groups of features that can be efficiently processed to built highly discriminative object models (Berg et al. 2005; Fergus et al. 2007; Leibe et al. 2004; Kouskouridas et al. 2013). According to Savarese and Fei-Fei (2007), the most important advantage of part-based frameworks stems from their ability to authorize condensed object models by computing the liaison between parts of an object from several views.

2 Method description

Initially, we collect images of large databases suitable for 3D object pose recovery containing several objects shot from varying perspectives. Our method utilizes the COIL-100 (Nayar et al. 1996) and CVL (Viksten et al. 2009) datasets, whilst incorporating synthetically rendered data of 3D models, available at http://www.evermotion.org/. At the present stage we hold a large number of training examples, which are images depicting various objects in arbitrary geometrical configurations along with the corresponding ground truth measurements stored as feature vector $\mathbf{y}_i \in \mathbb{R}^6$. The training set was chosen to be broad, since the proposed methodology does not employ an on-line training scheme and, thus, new training data cannot be introduced. The utilized datasets consists of images of objects captured under discrete intervals (e.g. of 5 degrees) in the 6-dimensional pose space, yet the regressor derives outputs in the continuous space. From every image in the training set we extract the respective SIFT (Lowe 2004) features that are post-processed by a homography-based RANSAC (Fischler and Bolles 1981) for outliers removal. Through the unsupervised clustering of the abstracted keypoints we define our bunch-based architecture. This is an additional feature representation subroutine, which builds new features considering both appearance and geometrical attributes of the training object. As a follow-up step the locations of each member of the architecture are provided as input to an ellipse fitting process in order to establish the initial pose manifolds. The latter represent the distances of each member of the part-based structure from the two foci of the fitting ellipse. To this point we would like to state that the main goal of the manifold modeling technique is to design a pose model that projects similar poses of different objects onto neighboring regions in the corresponding subspace. To this end, we consulted basic ellipse properties in the Euclidean Space implying that a 2D ellipse, opposed to circles or parabola, holds 5 DoFs (position, scale, shape and orientation) and stands for a five-dimensional manifold.

In the next step the initial manifolds are reprojected onto a highly representative and sparse subspace with large discrimination capabilities, via the minimization of the l_1 norm. The resulting training set is denoted as $\{\mathbf{r}_t, \mathbf{y}_t\}$ with $\mathbf{r}_t \in \mathcal{V}$ representing the constructed sparse manifolds (input vectors) and $\mathbf{y}_t \in \mathcal{Y}$ the respective ground truth measurements (output vectors). The process of building the training set is an iterative one that operates over several images of different objects. Our sparse pose manifold modeling method constructs $\mathbf{r} \in \mathcal{V}$ input vectors, taken into account during the process of learning of the Radial Basis Function regressor g, portraying the function $\phi(\mathbf{w})$. The latter represents the link that needs to be learnt, in order to establish the correct mapping between input vectors $\mathbf{r}_t \in \mathcal{V}$ and the modeled output vectors $\mathbf{y}_t = \mathbf{y}(\mathbf{r}_t; \phi(\mathbf{w})) \in \mathcal{Y}$, $\phi(\mathbf{w}) : \mathcal{V} \rightarrow \mathcal{Y}$. The ultimate goal of the proposed method is to achieve accurate estimations ($\mathbf{y}^*$) about the 3D pose for the unknown object o^* represented by $\mathbf{r}^*$ testing input vectors in order to facilitate the efficient manipulation of the object. A brief description of the method is illustrated by means of the block diagram of Fig. 2. The proposed sparse pose manifold method lays its foundations on four modules that are of fundamental importance and affect directly the discrimination capability of the established manifolds. These are: (i) a novel robust bunch-based structure; (ii) the ellipse fitting process and the initial manifold calculation; (iii) the reprojection of the resulting manifolds onto a more sparse and more representative subspace and, (iv) the training of the regressor. The above mentioned modules, all of which are discussed in detail later on in this section, provide contribution to the present state of the art either in their own or in combination and insofar as the authors are aware appear for the first time in the literature,

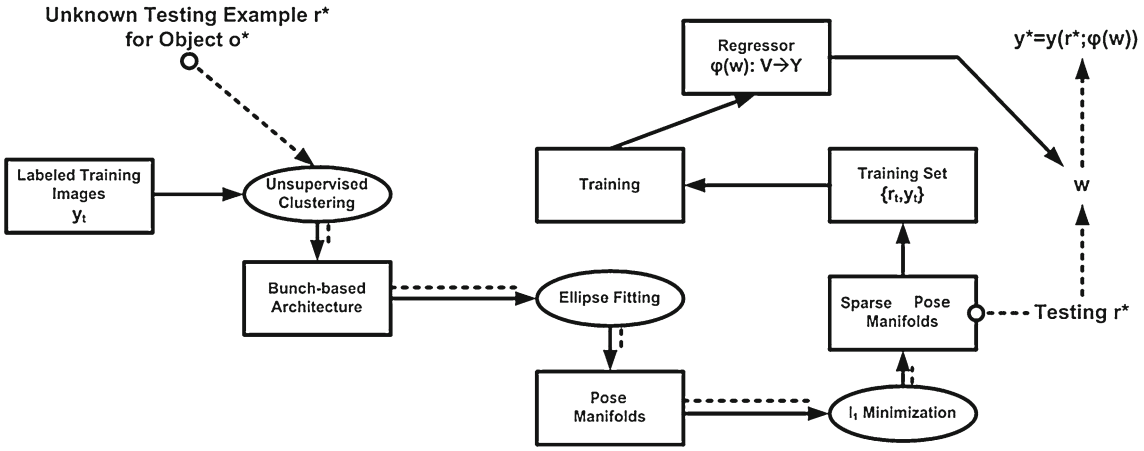


Fig. 2 The figure depicts the two distinguished phases of the proposed algorithm, namely the training (*continuous line*) and the testing (*dashed line*) ones. The main building blocks of the proposed method are: (a) the collection of labeled datasets dedicated to 3D pose estimation along with the corresponding ground truth measurement $\mathbf{y}_t$;(b) the module of building the bunch-based architecture through unsupervised clustering of the extracted visual cues; (c) the process of constructing initial manifolds via conforming an ellipse over the members of the bunch-based architecture; (d) the establishment of the highly representative subspace of sparse pose manifolds accomplished by l_1 norm minimization; (e) the training of the RBF-based regressor g, from *a priori* set $\{\mathbf{r}_t, \mathbf{y}_t\}$ with t training examples in order to estimate the continuous and invertible function $\phi(\mathbf{w}) : \mathcal{V} \rightarrow \mathcal{Y}$, with $\mathbf{w}$ representing the vector of adjustable parameters; (f) the process of executing a manipulation task for the unknown testing example $\mathbf{r}^*$ of object o^*

offering an elegant and efficient method for the manipulation of an arbitrarily placed object.

2.1 Bunch-based architecture

This module entails the subroutine of building a part-based formation that processes the abstracted visual cues of an object in an unsupervised manner. The most notable feature, constitutes the fact that our ultimate goal was to build a highly representative topology capable of modeling the visual data available in such a way so as to increase the generalization capacities of the resulting method. We employ a novel feature representation scheme to extract parts of objects that possess both appearance and geometrical properties of the target. Although our method shares common ground with the work of Mei et al. (2011), we consider the generalization capabilities of their technique being rather restricted due to their supervised-based feature extraction process. The latter extracts key-points through the realization of a Probability Density Function (PDF) associated with the joint distribution of appearance and geometry properties of the target.

Our bunch-based structure portrays appearance-based characteristics by extracting invariant features of high dimensionality with the SIFT descriptor (Lowe 2004) followed by RANSAC (Fischler and Bolles 1981) for outliers removal. Additionally, geometrical attributes of the object are aggregated by employing an unsupervised clustering technique over the matched SIFT key-points. Particularly, we treat the clustering problem from a Bayesian perspective by applying the Expectation Maximization (EM) algorithm to a Mixture Decomposition Scheme. (Please refer to Sect. 3.1 for additional information regarding the number of clusters). Analytical presentation of the mathematical formulation of the proposed part-based module is available in the Appendix. With a view to the reader's better understanding and in order to clarify the unsupervised clustering approach employed we introduce the respective pseudocode 1.

Algorithm 1 Calculate the positions $\boldsymbol{\theta}_\mathbf{k}$ of the γ clusters

1: **Inputs:** Locations of the ρ extracted SIFT features of the object o_i with pose q_j
2: **Initialize:** $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}(0)$, $\mathbf{P} \leftarrow \mathbf{P}(0)$, $\epsilon \leftarrow$very small number, $\tau = 0$
3: **while** $||\boldsymbol{\Theta}(\tau+1) - \boldsymbol{\Theta}(\tau)|| > \epsilon$ **do**
4: *Compute:* $P(\mathbf{b_k}|\mathbf{v}_\zeta; \boldsymbol{\Theta}(\tau)) \leftarrow (p(\mathbf{v}_\zeta|\mathbf{b_k}; \boldsymbol{\theta_k}(\tau))P_\mathbf{k}(\tau) \Big/ \sum_{k=1}^{\gamma}(p(\mathbf{v}_\zeta|\mathbf{b_k}; \boldsymbol{\theta}\mathbf{k}(\tau))P\mathbf{k}(\tau)$ with $\zeta = 1, \ldots, \rho$ and $k = 1, \ldots, \gamma$
5: *Find* $\boldsymbol{\theta_k}(\tau+1)$ *through the maximization of:* $\sum_{\zeta=1}^{\rho}\sum_{k=1}^{\gamma} P(\mathbf{b_k}|\mathbf{v}_\zeta; \boldsymbol{\Theta}(\tau)) \frac{\partial}{\partial \boldsymbol{\theta_k}} ln(p(\mathbf{v}_\zeta|\mathbf{b_k}; \boldsymbol{\theta_k}))$, w.r.t. $\boldsymbol{\theta}_\mathbf{k}$ for $k = 1, \ldots, \gamma$
6: *Set:* $P_k(\tau+1) \leftarrow \frac{1}{\rho}\sum_{\zeta=1}^{\rho} P(\mathbf{b_k}|\mathbf{v}_\zeta; \boldsymbol{\Theta}(\tau))$ with $k = 1, \ldots, \gamma$
7: $\tau = \tau + 1$
8: **end while**
9: **Outputs:** Locations of the γ clusters over the surface of the object

2.2 Establishing initial non-sparse manifolds

The goal of this module is to establish manifolds of low dimensionality capable of moderately distinguishing similar poses of different objects within the same pose subspace. Towards this end, we build a generalized model emphasizing

in encoding viewpoint-related data, invariant to object type and category. The resulting model should be able to project pose-related data onto the respective manifolds and to minimize intra-class variability. The proposed generalized module makes use of fundamental properties of the Euclidean space and its geometrical shapes. Essentially, our model focuses on fitting an ellipse on the members of the bunch-based structure of an object. It should be noted that fitting an ellipse on the contours of the Gaussians (estimated through the EM algorithm) resulted in a pose model that did not project similar poses of different object onto neighboring regions. Thus, our manifold modeling module employs the implicit conic equation of an ellipse with points (X, Y) in Cartesian coordinates, which takes the form $\psi(\Delta) = AX^2 + BXY + CY^2 + DX + EY + F = 0$. Here $\Delta = [A, B, C, D, E, F]$ corresponds to the vector of parameters to be estimated. Furthermore, we formulate the problem of the ellipse fitting as a combinatorial minimization problem of the following generalized least-squares cost function:

$$\begin{aligned} Q(\Delta) &= ||\hat{\mathbf{b}} - \psi||^2 \\ &= 1/\gamma \sum_{k=1}^{\gamma} ||AX^2 + BXY + CY^2 + DX + EY + F|| \\ &\quad + \lambda \sum_{N=1}^{5} \Delta_N^2 \end{aligned} \tag{1}$$

An object o_i with pose q is represented by the extracted clusters $\mathbf{b_k}, k = 1, \ldots, \gamma$, while N denotes the total number of parameters of the cost function Q. The regularized parameter λ is experimentally set to 0.1 and is summed over the first five elements of vector Δ excluding F that is regarded as bias. Additionally, the minimization of the cost function in Eq. 1 is accomplished via the PSO (Eberhart et al. 2001) tool, which stands for an evolutionary extrema estimation through the "social interaction" of a population of particles. Furthermore, we utilize PSO to find the particular particle-atom in the search space of 6 dimensions - as the dimensionality of the vector $\boldsymbol{\kappa}$- that holds the best score across all used generations. A more detailed analysis of the PSO falls beyond the scope of this paper; interested readers may refer to Eberhart et al. (2001) for additional information. We would like to note that utilizing the same regularization weight for different parameters, although it is definitely suboptimal in most of the cases, e.g. regularized linear regression, it is common place in PSO. Additionally, PSO was selected against other optimization techniques, e.g. gradient descent, Newton's method, due to the fact that PSO is computationally less expensive and converges faster. Moreover, in most of the cases gradient descent failed to converge to the global minimum, opposed to PSO.

2.3 Build sparse and representative pose manifolds

Given a sample $\boldsymbol{\omega} \in \mathbb{R}^m$ and a matrix $\mathfrak{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n] \in \mathbb{R}^{m \times n}$ where $\mathfrak{D}$ is an overcomplete dictionary with n elements, sparse representation attempts to derive a coefficient vector $\mathbf{s} \in \mathbb{R}^n$, possessing as many zero coefficients as possible and to simultaneously reconstruct $\boldsymbol{\omega}$ as a linear combination of $\mathfrak{D}$. This is formally expressed as:

$$\begin{aligned} &min_{\mathbf{s}} \quad ||\mathbf{s}||_0 \\ &\text{subject to} \quad \boldsymbol{\omega} = \mathfrak{D}\mathbf{s} \end{aligned} \tag{2}$$

where $||\mathbf{s}||_0$ denotes the l_0-norm, i.e. the number of non-zero elements of vector $\mathbf{s}$. However, solving Eq. 2 is an NP-hard problem, i.e. the most accurate method to derive a global solution is to exhaustively try out all possible subsets of $\mathbf{s}$. Nevertheless, the equivalence between l_0 and l_1 minimization problem has been shown, under the assumption that the solution $\mathbf{s}_0$ is sparse enough (Tsaig and Donoho 2006). Making one of those theoretical findings, the desired sparse solution $\mathbf{s}$ can be extracted via the minimization of the equivalent l_1 problem, which is solvable in polynomial time, i.e.:

$$\begin{aligned} &min_{\mathbf{s}} \quad ||\mathbf{s}||_1 \\ &\text{subject to} \quad \boldsymbol{\omega} = \mathfrak{D}\mathbf{s} \end{aligned} \tag{3}$$

Moreover, the strict equality constraint in Eq. 3 has been extended to an error-tolerant scheme, due to the existence of noise in real-life applications, leading to a more robust form:

$$\begin{aligned} &min_{\mathbf{s}} ||\mathbf{s}||_1 \\ &\text{subject to} \quad ||\boldsymbol{\omega} - \mathfrak{D}\mathbf{s}|| \leq \epsilon \end{aligned} \tag{4}$$

where $\epsilon \in \mathbb{R}$ is the error tolerance.

Once the data matrix $M = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_t] \in \mathbb{R}^{2*||\mathbf{b}_k|| \times t}$ has been extracted, where $2 * \mathbf{b}_j$ is the number of features and t is the number of samples, the l_1-norm is applied to every sample, expressing each one as a linear combination of the remaining data set. Moreover, the minimization criterion is enhanced by the addition of one extra constraint and is formulated as follows:

$$\begin{aligned} &min_{\mathbf{r}_i} ||\mathbf{r}_i||_1 \\ &\text{subject to} \quad \mathbf{m}_i = M\mathbf{r}_i, \quad i = 1, 2, \ldots, t \\ &\qquad\qquad 1 = \mathbf{1}^T \mathbf{r}_i \end{aligned} \tag{5}$$

For each $\mathbf{r}_i \in \mathbb{R}^t$ the $i - th$ value equals to zero, due to the fact that the corresponding $\mathbf{m}_i$ sample has been removed prior to the minimization. The elements of a $\mathbf{r}_i$ vector denote the degree of participation of the remaining data set in order $(\mathbf{r}_i)$ to be reconstructed. The reconstruction weight $\mathbf{r}_i$ provides significant geometric capabilities, such as invariance to rotations and scalings, while the new imposed constraint attaches consistency to translations. Through this process we

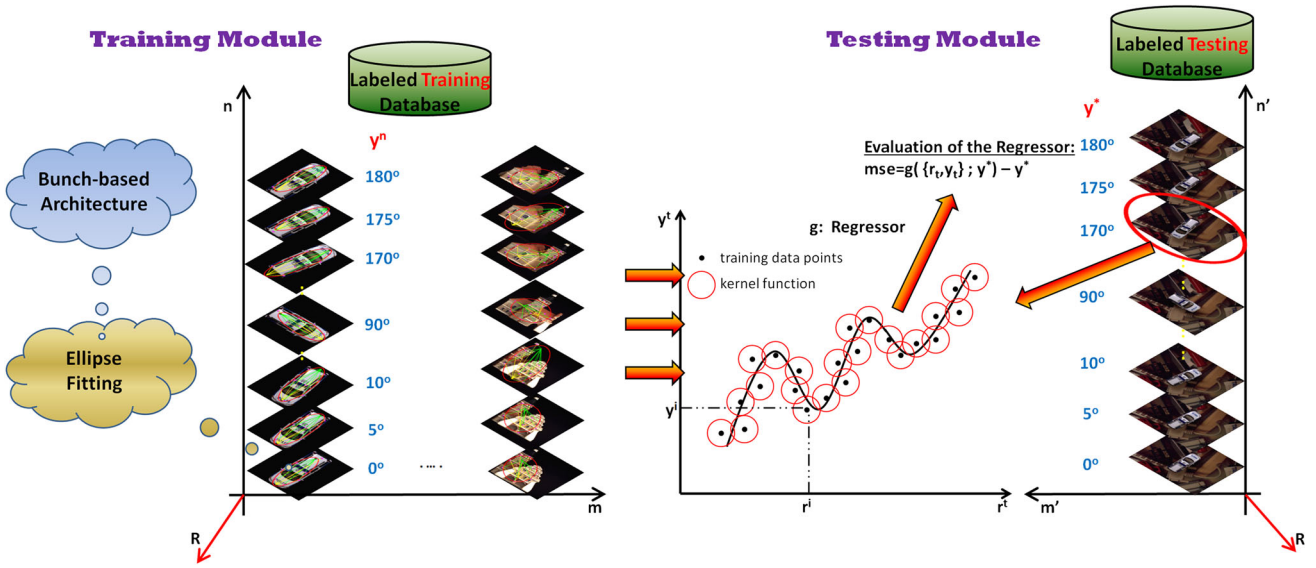


Fig. 3 The process of building the training set $\{\mathbf{r}_t, \mathbf{y}_t\}$ used for the 3D pose estimation module incorporates the division of the labeled datasets into the training and the testing subsets, respectively. The built training set is fed to the Radial Basis Function-based regressor g in order to produce an accurate mapping between the input space $\mathcal{V}$ and the output space $\mathcal{Y}$

reproject the initial pose manifolds onto more compact, highly representative subspaces.

2.4 Training phase

As described above, the members of the bunch-based architecture (clusters $\mathbf{b_k}$) are fed to the ellipse fitting module for the purpose of efficient manifold modeling. The aforementioned procedure proceeds with iteration over several images of numerous objects. Along with this repeated process there is a need to cope with the 2D-2D feature correspondence problem to recall patterns that hold highly distinguishable data. We have gathered a large collection of object images shot every 5^0 accompanied by the corresponding ground truth 6 DoF pose measurements stored in the vector $\mathbf{y}_t$. Figure 3 is drawn with the view to illustrate the process of building the training set $\{\mathbf{r}_t, \mathbf{y}_t\}$, the task of sparse pose manifolds modeling and network simulation.

3 Experimental results

In the following section we present both qualitative and quantitative results, as obtained by the proposed method. It becomes apparent that the effectiveness of the method presented in this paper relies on its efficiency to build highly representative and sparse manifolds with the view to provide accurate 6 DoF measurements related to geometrical configurations of objects in the 3D space. Towards this end we compared the efficiency of the proposed framework with other state-of-the-art approaches used in the solution of the 3D object pose recovering problem. More specifically, the methods chosen:

- Incorporate a part-based module for 3D pose estimation like that in Hinterstoisser et al. (2007) (abbreviated as *N3M*)
- Emphasize on manifold modeling for 3D pose recovering like that in Mei et al. (2011) (abbreviated as *SMM*)
- Utilize a common least squares-based benchmark technique (Lowe 1999); also adopted in proposed frameworks more recently (Ma et al. 2011) (abbreviated as *LS-based*)
- Encompass a conventional dimensionality reduction scheme as presented in Yuan and Niemann (2001) (abbreviated as *PCA-based*).

At this point, we wish to point out the critical lack of databases containing images of objects in shifting 3D geometrical configurations, together with their corresponding ground truth measurements. Here, we make use of the only available datasets viz. COIL-100 (Nayar et al. 1996) and CVL(Viksten et al. 2009), whilst we address the issue of the limited amount of 3D datasets by expanding our training set with artificially rendered images for a collection of objects available at http://www.evermotion.org/. The training set is essentially comprised of object images from the COIL-100 (Nayar et al. 1996), CVL (Viksten et al. 2009) (images of uncluttered background) and Evermotion (2012) databases. Moreover, in order to derive accurate and representative results we performed a leave-one-out notion, with

respect to the classes. Additionally, the efficiency of the proposed technique is further assessed by comparison to real images of objects in cluttered environments (Viksten et al. 2009).

3.1 Number of clusters $\mathbf{b}_k$ and size of the training set $\{\mathbf{r}_t, \mathbf{y}_t\}$

The constructed sparse pose manifolds, fed to the RBF-based regressor to produce the accurate input-output space mapping, are of $2*||\mathbf{b}_k||$ dimensionality (with $\mathbf{b}_k$ representing the number of the members of the bunch-based structure extracted via unsupervised clustering). According to the literature, there are two ways to choose the correct number of clusters: a) the "elbow" method, where the right number of clusters is considered producing the maximum discrepancy between two neighboring values of the cost function; b) by evaluating the clustering process based on a downstream purpose (in our case the 6 DoF object pose estimation). We have adopted the second approach since we wish to solve the object pose estimation problem, rather than the efficient clustering one. Accordingly, and after exhaustive testing, the number of clusters was set to 8, as it exhibits low generalization error at high performance levels. We should further point out that our method is not limited necessarily to the selection of SIFT or the Mixture Decomposition Scheme. It is, nevertheless, true that the aforementioned algorithms have shown to be more effective in representing both appearance and geometry of the objects. Moreover, since affine changes in illumination may affect the performance of the descriptor, each feature vector is normalized to unit length for the sake of preserving invariance under such circumstances. Accurate 3D target pose measurements are provided through the simulation of a regressor whose input vectors are preprocessed by a RBF kernel. The same principle applies to this feature of the proposed algorithm; in the sense that our system does not depend on the selection of the kernel, even if the RBF provided more accurate results.

Our labeled training set $\{\mathbf{r}_t, \mathbf{y}_t\}$ is built through an iterative procedure that constructs t training examples by means of the architecture presented in Sect. 2. Input vectors $\mathbf{r}_t$ of $2*||\mathbf{b}_k||$ dimensionality are accompanied with the respective 6-dimensional ground truth estimations $\mathbf{y}_t$ corresponding to known object poses. Objects used for training are either placed on a turntable and shot every 5^0 with changing camera poses (Viksten et al. 2009) or artificially rendered (http://www.evermotion.org/) under varying viewpoints through Blender (2011). 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, without the addition of noisy vectors or others caused by partially occluded objects, is $[2 * \mathbf{b}_k \times 100,000]$, that is 1,000 images/object. Ground truth measurements $\mathbf{y}_t$ are determined by storing the camera's rotational and translational parameters through a predefined moving path in the 3D space (6 DoF) and with a given time step. With the addition of noisy input vectors and partial occlusions-related ones, the size of the training set $\{\mathbf{r}_t, \mathbf{y}_t\}$ is $[2 * \mathbf{b}_k \times 250,000]$. This results in large computational burden for even a state-of-the-art workstation; typically requiring approximately 200 h on the expanded version of the training set. Preliminary experimental results for both un-occluded objects and partially occluded targets are depicted in Fig. 4.

3.2 Dealing with occlusions

One of the most frequent problems in computer vision is the partial occlusion of objects with which contemporary frameworks are severely restricted in coping with. We consider this issue by introducing (a) noisy input vectors and (b) images of objects with artificially generated partial occlusions to various degrees. The percentage of obstruction lays in the range of [0–60] with a black rectangle of random size being arbitrarily overlaid on the surface of the training object. As shown in Bishop (2006), training with noise approximates Tikhonov's regularization theorem, whilst bootstrapping the performance of the learning procedure. In the Appendix we demonstrate how the training with noisy input vectors is adjusted, adequately in our case. Figure 4 illustrates the visual outcome of the proposed 3D object pose estimation module for non-occluded and partially occluded objects. We deal with the cluttered background by utilizing a contemporary image segmentation technique, such as that presented in Shi and Malik (2000). In order to evaluate the performance of the proposed 3D pose estimation approach under various degrees of partial occlusions we make use of the criterion stated in Hinterstoisser et al. (2007). According to this particular metric, a measurement of the 3D pose of a target is taken as accurate for all the cases where the sum of error of the rotation parameters is less than 5^0. As depicted in Fig. 5(a), our sparse pose manifold modeling approach is shown to be more tolerant to partial occlusions (spanning 0 to 95 percentage of coverage) as opposed to *N3M* (Hinterstoisser et al. 2007), *SMM* (Mei et al. 2011), *LS-based* (Ma et al. 2011) and *PCA-based* (Yuan and Niemann 2001). We have further tested the accuracy of our method, since both our training set and those of other related works contain images of objects shot every 5^0 that might have resulted in overfiting the built models. Figure 5(b) depicts the respective comparative results for a permissible error of 3^0, whereas both experimental sets of Fig. 5 show results obtained from simulated networks with over 250 testing examples.

Fig. 4 The visual outcome of the proposed approach for un-occluded and partially occluded (under varying percentages) objects in cluttered backgrounds

3.3 Accuracy of 6 DoF pose estimation and object manipulation

In this section we demonstrate the accuracy of the *Sparse Pose Manifolds* method for the particular tasks of 6 DoF pose recovering and object manipulation. With a view our method to be comparable to other state-of-the-art methodologies, the experimental setup for eye-in-hand performance testing shares common spirit with the one presented in Ma et al. (2011). However, we possess a 5 DoF robotic arm that does not allow a complete evaluation of the 6 DoF estimation of our method. We have addressed this issue by introducing rendered images of different objects captured by a moving camera in an artificial environment (http://www.blender.org/). The pose parameters of the camera are considered to be the groundtruth measurements. Furthermore, in order to demonstrate the generalization capabilities of our method we utilize unknown objects that belong to 6 general object classes, i.e. box-shaped objects (e.g. a Rubik's cube), cars, motorcycles, balls, 4-legged animals and cups. Besides these 6 object classes, the proposed method is able to generalize to handle arbitrarily shaped objects. Ground truth measurements correspond to the pose parameters of a moving camera fixated to the test object laying at a predefined position in the 3D space. Figures 6 and 7 illustrate the efficiency of our method for all the test objects of the 6 different object classes. Thereupon, the proposed method is shown to exhibit high efficiency levels in recovering the 3 DoF translation parameters since they

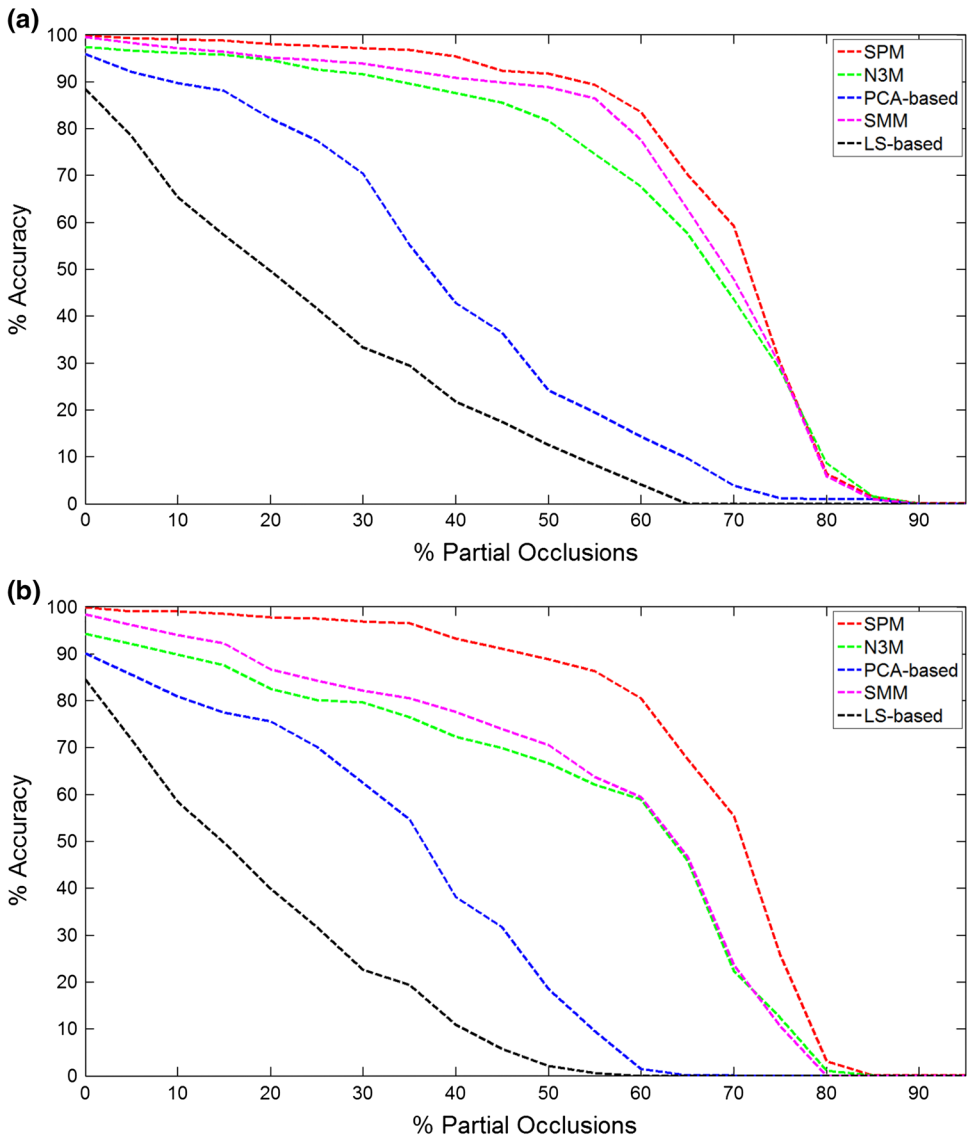


Fig. 5 Comparative evaluation of the tolerance of the proposed method under various degrees of partial occlusions for a permissible error of 5^0 (**a**) and 3^0 (**b**), respectively

match, almost exactly, the ground truth measurements provided through rendering. Moreover, the estimated rotational parameters for all the objects under test only slightly differ from the ground truth.

The proposed method is capable of providing accurate 6 DoF pose estimations regardless of the class of the object (see Figs. 6 and 7). From the detailed study of the experimental simulations several conclusions are drawn about the underlying capacities of the *Sparse Pose Manifolds*. In particular, objects belonging to the classes of cup and 4-legged animal, are seen to cause larger oscillations in the estimated rotational parameters, leading to a slightly decreased performance of the object manipulation module as a result. In essence, the more regular the shape of an object the higher the performance of the 3D pose recovering and grasping modules. More specifically, it is noted that ellipses with lower eccentricities lead to more distinguishable manifolds, while circle-like extracted ellipses lead to low dimensional data that result in notable degradations in the performance of the 6 DoF pose estimation method. As a result, prismatic objects, such as cars and motorcycles evinced higher grasping accomplishment frequencies as contrasted to balls, 4-legged animals and cups. As the estimation of the pose of any object relies on the ellipse fitting module, it is straightforward to claim that the designed pose model is feature-sensitive rather than shape-based. This particular attribute of our method facilitates the efficient addressing of issues raised by symmetrical objects, e.g. a ball. Table 1 depicts the condensed results after executing 250 grasping operations for 6 different objects. Although we possess a 5 DoF robotic arm, we tackle the non-holonomic problem by positioning the objects on frames perpendicular to the working table, whilst the respective response of the network was set to zero. Moreover, a manipulation operation is considered as a successful one, in cases where the robot arm flawlessly grasps the testing object. We would like to note that the mechanical attributes of the gripper enabled

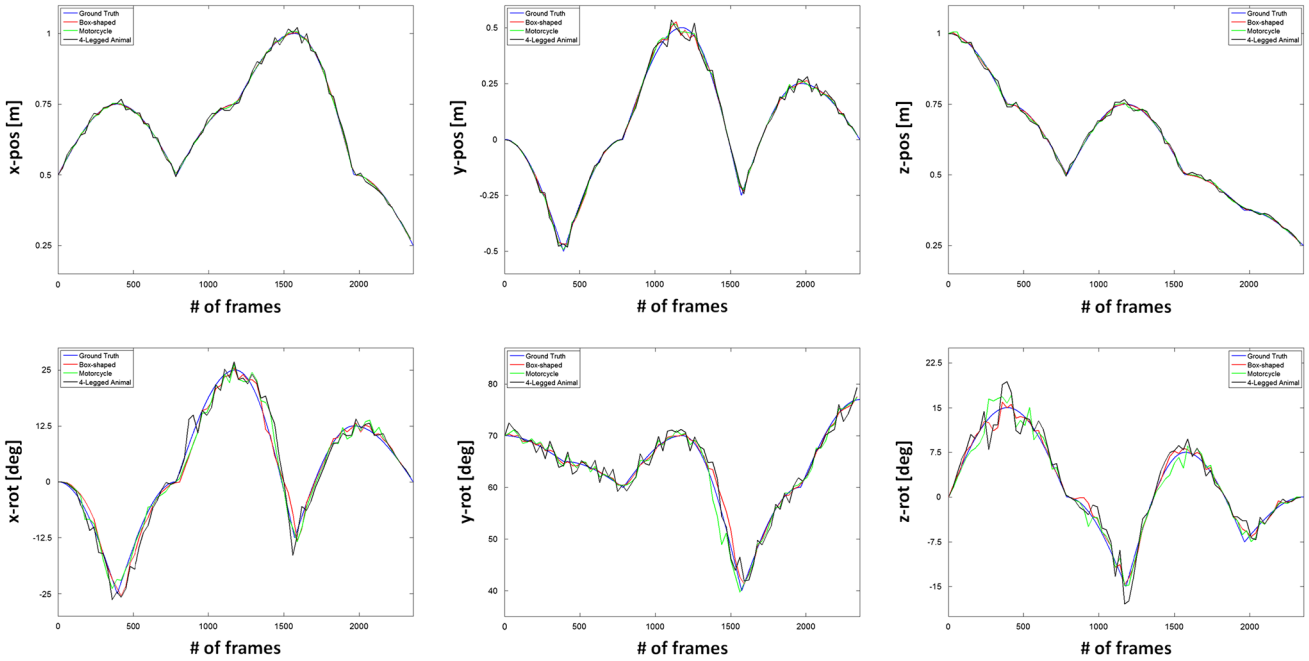


Fig. 6 6 DoF object pose estimation for a box-shaped object, a motorcycle and a 4-legged animal, along with the corresponding ground truth measurements

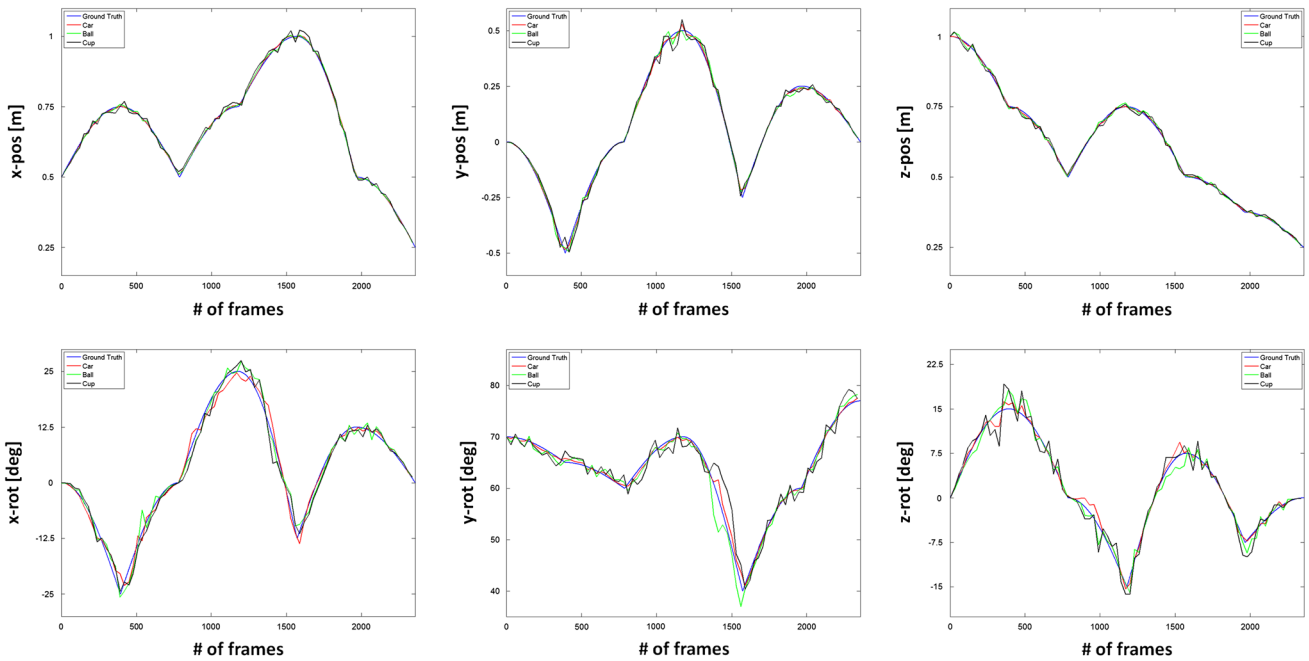


Fig. 7 6 DoF object pose estimation for a car, a ball and a cup, along with the corresponding ground truth measurements

the manipulation of any object belonging to the respective classes.

The proposed methodology was compared to the classic KNN classification algorithm. The training matrix for the KNN was the same as the one used for the regressor and the corresponding labels are the poses (in degrees) of training samples. Concerning whether a pose of a test sample has been accurately classified, we provided an error tolerance of ± 5 degrees from the ground truth. The KNN was tested using 1 and 3 number of neighbors, i.e K=1 and K=3. The results are summarized in Table 2. The accuracy of the KNN in comparison with the regressor in the proposed technique is significantly inferior, for the aforementioned reasons. Besides the reduction of the success rate, the standard deviation was

Table 1 This summary table illustrates the accuracy along with the corresponding standard deviations (σ) of the proposed method in accomplishing grasping operations with respect to multiple objects belonging to 6 different classes

Object class	Success rate %	Std (σ)	Mean eccentricity	Std (σ)
Box-shaped	97.2	0.9	0.31	0.09
Car	95.6	1.1	0.44	0.12
Motorcycle	92.8	1.6	0.61	0.11
Ball	88.4	2.5	0.76	0.08
4-legged animal	87.9	2.9	0.84	0.09
Cup	86.6	3.3	0.87	0.07

As it can be seen, the success rate is closely related to the mean eccentricity of the extracted ellipses, since smaller eccentricity values imply higher manipulation rates

Table 2 Summarizing results for the KNN classification experiments

Object class	Success rate % (K=1)	Std (σ) (K=1)	Success rate % (K=3)	Std (σ) (K=3)
Box shaped	94.7	2.6	94.2	2.3
Car	91.5	3.2	94.5	3.0
Motorcycle	92.0	1.8	95.5	1.1
Ball	86.2	3.4	86.1	2.5
4-legged animal	83.3	3.7	83.8	2.9
Cup	82.6	3.4	82.8	3.2

increased dramatically. There are object classes, such as the box shaped and the car where the standard deviation was over-doubled. The comparison between the number of neighbors, suggest that for K=3 the system improved its success rate in most cases, yet for two object classes, box shaped and ball, the usage of K=1 provided slightly better results. Moreover, for K=3 the standard deviation was improved for five out of six classes, indicating that for number of neighbors K=3 the system provide more stable results.

3.4 Real-time execution and testing with real objects

The calibration procedure set the camera's extrinsic parameters corresponding to its pose in the 3D working space, in addition to its holding of information of its position relative to end effector's coordinate system and the frame of the robot itself. The experiments ran on an intel i5 2.4 GHz dual core processor with 8 GB RAM. The resolution of images was 640×480 pixels, while the frame rate was 5 Hz. The most time demanding module of the proposed method (apart from the training session which is performed off-line) is that of extracting SIFT features. To achieve real-time execution, SIFT is applied over the complete test image only for the first frame, with the view to obtain initial set of SIFT features. Then, at each subsequent time step, SIFT extraction is limited to a smaller region of interest (ROI) within the image containing features whose distance from their mean is smaller than 3 times (set experimentally) their standard deviation. We have further evaluated the performance of our method, in several experimental tests that involve either eye-in-hand or eye-to-hand camera configurations with the use of realobjects. In the videos accompanying the paper in hand[1], it can be seen: (a) how our method is capable of efficiently guiding a robotic gripper to grasp an unknown object (eye-to-hand version) and (b) the performance of the 6 DoF pose estimation module (eye-in-hand version). Figures 8 and 9 present in a more illustrative form the execution stages of the proposed method, in its eye-to-hand version, from the process of ellipse fitting to the efficient accomplishment of manipulation tasks for a car and a box, respectively. The lighting conditions were kept controlled, as the main goal in the conducted experiments was to prove that sparse pose manifolds are feasible and can provide efficient solutions to the pose estimation problem. In a more demanding scenario, the proposed algorithm can be enriched with illumination compensation capabilities (Vonikakis et al. 2013).

4 Discussion

In this paper we present a novel solution to the automatic manipulation of unknown objects based on the first integrated research attempt to formulate *Sparse Pose Manifolds*. The key ideas underlying the proposed method are that: a) different objects viewed under the same perspective share identical poses, which can be efficiently projected onto a well-defined and highly distinguishable subspace; and b) the reprojection of initial manifolds onto sparse and highly representative subspaces lead to increased generalization capabilities of the method. Towards this end, we propose an integrated image understanding architecture that imposes upon a bunch-based structure, involving the unsupervised clustering of the extracted features, whilst encapsulating both appearance and geometrical attributes of the objects. Through the process of ellipse fitting we build initial pose manifolds of low dimensionality, therefore, minimizing the complexity of the system and avoiding the usage of conventional dimensionality reduction schemes. In addition, we post-process the established pose manifolds via l_1 norm minimization to build sparse and compact input vectors characterized by large discrimination and generalization capacities. We believe our work makes the following contributions: a) we show that there are

[1] Grasping a pliers:http://www.youtube.com/watch?v=J_gpPu6ZYQQ, Grasping a box-shaped object:http://www.youtube.com/watch?v=cOF0RdeJ6Zg, pose estimation of a car: http://www.youtube.com/watch?v=CSoFMk48DmM, pose estimation of a box-shaped object http://www.youtube.com/watch?v=RTe6usXm9qs.

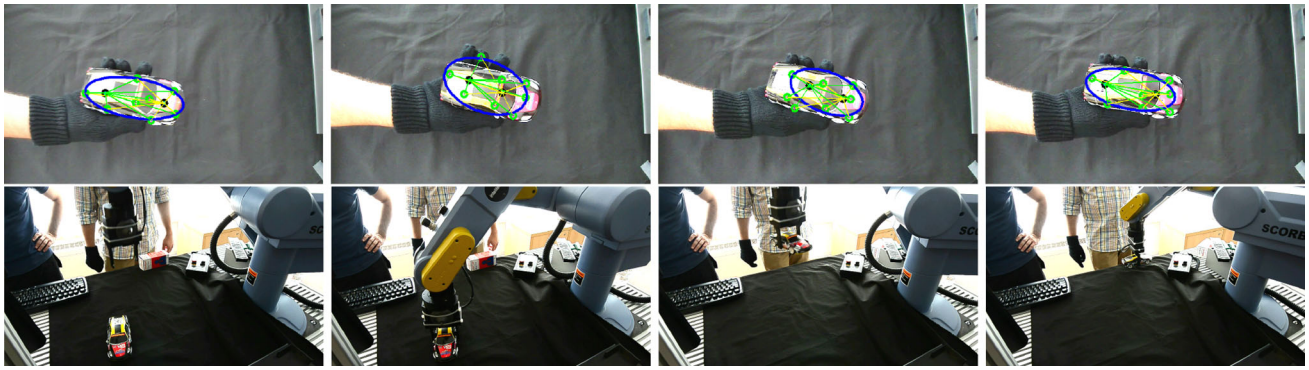


Fig. 8 *Upper*: Images obtained through the eye-to-hand camera are fed to the ellipse fitting process to provide the corresponding results. Here, the *black* circles represent the two foci of the ellipse while the *green* ones stand for the members of the bunch-based architecture. *Lower*: After the establishment of the initial pose manifolds and their reprojection onto a more sparse and highly representative subspace, the robotic arm is adequately configured to grasp an unknown test object (e.g. car)

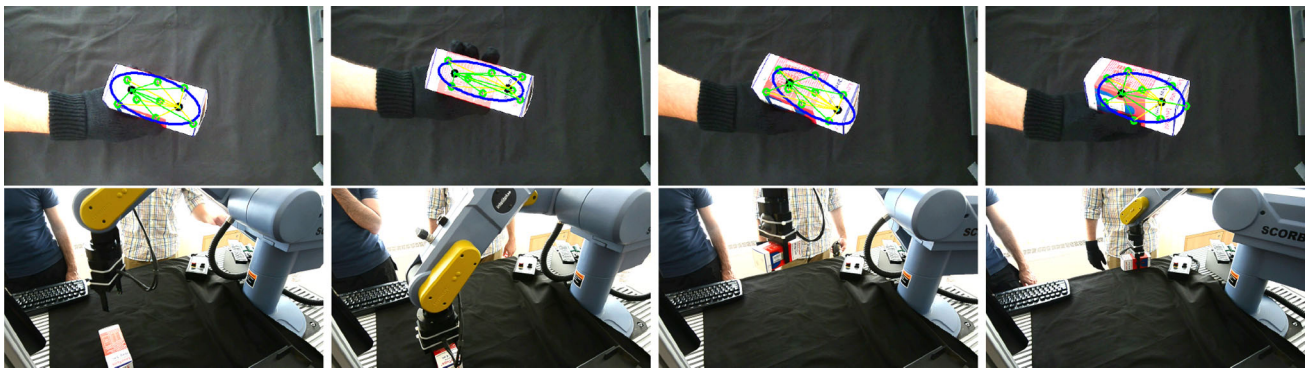


Fig. 9 *Upper*: Regardless of the class of the unknown test object the ellipse fitting process is performed every 5 frames in order to produce low dimensional sparse input vectors. *Lower*: The latter are fed to the regressor that provides accurate estimations about the pose of the unknown test object (e.g. box). These measurements are then transfered to the controller of the robot that, in turn, adjusts its 5 DoF joints to perform a grasping task

indeed manifolds that can be sufficiently modeled through a training session with several objects under varying poses; b) the proposed bunch-based architecture extracts new features that enjoy both appearance and geometrical attributes of the objects, whilst demanding less supervision during training; c) through the minimization of the l_1 norm we establish sparse and representative pose manifolds that are fed to a RBF-based regressor to recover the mapping between input and output spaces.

We can furthermore claim that our work is shown to be more tolerant to partial occlusions as compared to other related methods, primarily due to the following reasons: Our bunch-based architecture extracts features that enjoy both appearance and topological attributes, in contrast to the *N3M* method, for example, requiring large amount of inter-part connections for accurate measurements. Regarding the *SMM* method, it is noted that it abstracts parts of objects in a supervised manner through the realization of the PDF associated to the joint distribution of appearance and geometry properties of the object. However, the main drawback of *SMM* is that it is very sensitive to partial occlusions due to the fact that the extracted parts correspond to specific regions on the surface of the object (e.g. the tire of a car) notwithstanding the limited dataset used for training (as it contains only cars). Moreover, the *SMM* method relies on a statistical manifold modeling approach that is based on the operations of expansion and alignment in order to minimize the labeling ambiguity. The authors of *SMM* utilize video sequences in place of still images, hence, decreasing the reliability of the ground truth measurements. These operations are used in order to facilitate the efficient categorization of testing objects into the respective classes (which hold data of poses that are similar to those used for training). In contrast, the proposed *SPM* modeling technique establishes manifolds that, through the unsupervised clustering subroutine and the ellipse fitting procedure, hold highly distinguishable data that can be easily classified into the respective classes. Furthermore, our method is also less affected by partial occlusions as compared to the

LS-based and *PCA-based* methods primarily due to the linear mapping between input and output space and, secondly, due to the high dimensional input vectors of limited discrimination capacities. More conclusively, experimental results provide evidence of low generalization error rendering thus justification to our theoretical claims. It should also be noted that, box-shaped objects (with low eccentricity values) as opposed to objects belonging to the classes of 4-legged animals, cups and balls (with higher eccentricity values), achieve higher rates of success in the particular task of performing a grasp on unknown targets. Although sphere-like objects are the most regular objects in the real world, their projection onto a 2D image forms a circle that possesses fewer DoFs than an ellipse, which leads to singular configurations. Moreover, the extracted features of targets belonging to the classes of 4-legged animals, cups and balls are closer to a circle rather than an ellipse and exhibit slightly reduced rates of success in object manipulation tasks, as a result. Finally, with regards to future work, we aim to concentrate on designing and implementing an advanced manifold modeling structure that lays its foundations in the attributes of a 3D ellipse fitted over 3D features obtained from an RGB-D depth camera. Intuitively, such a new architecture accompanied by an exhaustive training session with numerous examples should represent a new stream of solutions to the object manipulation problem.

Acknowledgments The authors would like to thank Nikolaos Metaxas-Mariatos for his help in conducting the experimental validation of the proposed method.

Appendix

Formulation of bunch-based architecture

Given an image of an object with certain pose, we first extract the 2D locations of the ρ SIFT keypoints denoted as $\mathbf{v}_\zeta \in \mathbb{R}^2$. Then we perform clustering over the locations of the extracted interest points (input vectors $\mathbf{v}_\zeta$, $\zeta = 1, 2, \ldots, \rho$) in order to account for the topological attributes of the object. Supposing there are γ clusters denoted as $\mathbf{b_k}$, $k = 1, 2, \ldots \gamma$, we consider basic Bayesian rules noting that a vector $\mathbf{v}_\zeta$ belongs to a cluster $\mathbf{b_k}$ if $P(\mathbf{b_k}|\mathbf{v}_\zeta) > P(\mathbf{b}_\zeta|\mathbf{v_j})$, $\zeta, k = 1, 2, \ldots, \gamma$, $\zeta \neq j$. The expectation of the unknown parameters conditioned on the current estimates $\boldsymbol{\Theta}(\tau)$ (τ denotes the iteration step) and the training samples (E-step of the EM algorithm) are:

$$J(\boldsymbol{\Theta};\ \boldsymbol{\Theta}(\tau)) = E\Big[\sum_{i=1}^{\rho} ln(p(\mathbf{v}_\zeta|\mathbf{b_k};\boldsymbol{\theta})P_\mathbf{k})\Big] = \sum_{\zeta=1}^{\rho}\sum_{k=1}^{\gamma} P(\mathbf{b_k}|\mathbf{v}_\zeta;\boldsymbol{\Theta}(\tau))ln(p(\mathbf{v}_\zeta|\mathbf{b_k};\boldsymbol{\theta})P_\mathbf{k}) \quad (6)$$

with $\mathbf{P}_{1\times\gamma} = [P_1, \ldots, P_\gamma]^T$ denoting the *a priori* probability of the respective clusters, $\widehat{\boldsymbol{\theta}_{2\times\gamma}} = [\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}\gamma^T]^T$ corresponding to the $\boldsymbol{\theta_k}$ vector of parameters for the $k-th$ cluster and $\boldsymbol{\Theta}_{3\times\gamma} = [\hat{\boldsymbol{\theta}^T}, \mathbf{P}^T]^T$. According to M-step of the EM algorithm, the parameters of the γ clusters in the respective subspace are estimated through the maximization of the expectation:

$$\boldsymbol{\Theta}(\tau+1) = \arg\max_{\boldsymbol{\Theta}} J(\boldsymbol{\Theta};\boldsymbol{\Theta}(\tau)) \quad (7)$$

resulting in

$$\sum_{\zeta=1}^{\rho}\sum_{k=1}^{\gamma} P(\mathbf{b_k}|\mathbf{v}_\zeta;\boldsymbol{\Theta}(\tau))\frac{\partial}{\partial\boldsymbol{\theta_k}} ln(p(\mathbf{v}_\zeta|\mathbf{b_k};\boldsymbol{\theta_k})) = 0 \quad (8)$$

while maximization with respect to the *a priori* probability $\mathbf{P}$ gives:

$$P_k = \frac{1}{\rho}\sum_{\zeta=1}^{\rho} P(\mathbf{b_k}|\mathbf{v}_\zeta;\boldsymbol{\Theta}(\tau)) \qquad \text{with } k = 1, \ldots, \gamma \quad (9)$$

It is apparent that the optimization of Eq. 8 with respect to $\mathbf{P}$ stands for a constraint maximization problem that has to obey to $P_k \geq 0$, $k = 1, \ldots, \gamma$ and $\sum_{k=1}^{\gamma} P_k = 1$. We revise the Lagrangian theory that states:

> *Given a function $f(x)$ to be optimized subject to several constraints built the corresponding Lagrangian function as $\mathcal{L}(x, \lambda) = f(x) - \sum \lambda f(x)$.*

Following on from the above statement, we denote the respective (to Eq. 6) Lagrangian function as:

$$\mathcal{J}(\mathbf{P}, \lambda) = J(\boldsymbol{\Theta};\boldsymbol{\Theta}(\tau)) - \lambda\Big(\sum_{k=1}^{\gamma} P_k - 1\Big)$$

We obtain λ and P_k though:

$$\partial\frac{\mathcal{J}(\mathbf{P},\lambda)}{\partial P_k} = 0 \Rightarrow$$

$$\partial\frac{\Big(\sum_{\zeta=1}^{\rho}\sum_{k=1}^{\gamma} P(\mathbf{b_k}|\mathbf{v}_\zeta;\boldsymbol{\Theta}(\tau))ln(p(\mathbf{v}_\zeta|\mathbf{b_k};\boldsymbol{\theta})P_\mathbf{k})\Big)}{\partial P_k} -$$

$$-\frac{\lambda(\sum_{k=1}^{\gamma} P_k - 1}{\partial P_k} = 0 \Rightarrow$$

$$P_k = \frac{1}{\lambda}\sum_{\zeta=1}^{\lambda} P(\mathbf{b_k}|\mathbf{v}_\zeta;\boldsymbol{\Theta}(\tau))$$

Since $\sum_{k=1}^{\gamma} P_k = 1$, we can derive that $\lambda = \rho$

resulting in the final $apriori$ probability of the $k-th$ cluster of Eq. 9:

$$P_k = \frac{1}{\rho}\sum_{\zeta=1}^{\rho} P(\mathbf{b_k}|\mathbf{v}_\zeta; \boldsymbol{\Theta}(\tau)) \qquad \text{with } k = 1, \ldots, \gamma$$

Training with noise

The performance of the proposed regressor-based 3D pose estimation module is bootstrapped by adding noise to the input vectors fed to the RBF-kernel during training. In the following passage we present a slightly modified version of the Tikhonov regularization theorem as adjusted to the needs of our case. In cases where the inputs do not contain noise and the size $\mathbf{t}$ of the training dataset tends to infinity, the error function containing the joint distributions $p(\mathbf{y}_\lambda, \mathbf{r})$ (of the desired values for the network output $\mathbf{g}_\lambda$) assumes the form:

$$\begin{aligned} E &= \lim_{\mathbf{t}\to\infty} \frac{1}{2\mathbf{t}} \sum_{k=1}^{\mathbf{t}} \sum_{\lambda} \{\mathbf{g}_\lambda(\mathbf{r_k}; \mathbf{w}) - \mathbf{y}_{\lambda\mathbf{k}}\}^2 \\ &= \frac{1}{2}\sum_m \int\int \{\mathbf{g}_\lambda(\mathbf{r_k}; \mathbf{w}) - \mathbf{y}_{\lambda\mathbf{k}}\}^2 p(\mathbf{y}_\lambda, \mathbf{r}) d\mathbf{y}_\lambda d\mathbf{r} \\ &= \frac{1}{2}\sum_m \int\int \{\mathbf{g}_\lambda(\mathbf{r_k}; \mathbf{w}) - \mathbf{y}_{\lambda\mathbf{k}}\}^2 p(\mathbf{y}_\lambda|\mathbf{r}) p(\mathbf{r}) d\mathbf{y}_\lambda d\mathbf{r} \end{aligned}$$

Let $\boldsymbol{\eta}$ be a random vector describing the input data with probability distribution $p(\boldsymbol{\eta})$. In most of the cases, noise distribution is chosen to have zero mean ($\int \eta_i p(\boldsymbol{\eta}) d\boldsymbol{\eta} = 0$) and to be uncorrelated ($\int \eta_i \eta_j p(\boldsymbol{\eta}) d\boldsymbol{\eta} = \text{variance}\sigma_{ij}$). In cases where each input data point contains additional noise and is repeated infinite times, the error function over the expanded data can be written as:

$$\begin{aligned} \widetilde{E} &= \frac{1}{2}\sum_m \int\int\int \{\mathbf{g}_\lambda((\mathbf{r_t}; \mathbf{w}) + \boldsymbol{\eta}) - \mathbf{y}_{\lambda\mathbf{k}}\}^2 \\ &\quad p(\mathbf{y}_\lambda \mid \mathbf{r}) p(\mathbf{r}) p(\boldsymbol{\eta}) d\mathbf{y}_\lambda d\mathbf{r} d\boldsymbol{\eta} \end{aligned}$$

Expanding the network function as a Taylor series in powers of η produces:

$$\begin{aligned} \mathbf{g}_\lambda((\mathbf{r_t}; \mathbf{w}) + \boldsymbol{\eta}) &= \mathbf{g}_\lambda(\mathbf{r_t}; \mathbf{w}) + \sum_i \eta_i \frac{\partial \mathbf{g}_\lambda}{\partial \mathbf{r}_i}\bigg|_{\eta=0} + \\ &+ \frac{1}{2}\sum_i \sum_j \eta_i \eta_j \frac{\partial^2 \mathbf{g}_\lambda}{\partial \mathbf{r}_i \partial \mathbf{r}_j}\bigg|_{\eta=0} + \mathcal{O}(\eta^3) \end{aligned}$$

By substituting the Taylor series expansion into the error function we obtain the following form of regularization term that governs the Tikhonov regularization:

$$\widetilde{E} = E + variance \times \Omega$$

with

$$\begin{aligned} \Omega = \frac{1}{2}\sum_m \sum_i \int\int \left\{ \left(\frac{\partial \mathbf{g}_\lambda}{\partial \mathbf{r}_i}\right)^2 + \frac{1}{2}\{\mathbf{g}_\lambda(\mathbf{r}) - \mathbf{y}_\lambda\} \frac{\partial^2 \mathbf{g}_\lambda}{\partial \mathbf{r}_i^2} \right\} \\ p(\mathbf{y}_\lambda|\mathbf{r}) p(\mathbf{r}) d\mathbf{r} d\mathbf{y}_\lambda \end{aligned}$$

References

Agrawal, A., Sun, Y., Barnwell, J., & Raskar, R. (2010). Vision-guided robot system for picking objects by casting shadows. *IJRR*, *29*, 155–173.

Andreopoulos, A., Tsotsos, J. (2009). A theory of active object localization. *ICCV* (pp. 903–910).

Baudat, G., & Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation*, *12*, 2385–2404.

Ben Amor, H., Kroemer, O., Hillenbrand, U., Neumann, G., Peters, J. (2012). Generalization of human grasping for multi-fingered robot hands. In *Intelligent Robots and Systems (IROS)*, 2012 IEEE/RSJ International Conference on. (pp. 2043–2050).

Berg, A., Berg, T., & Malik, J. (2005). Shape matching and object recognition using low distortion correspondences. *CVPR*, *1*, 26–33.

Bishop, C. (2006). *Pattern recognition and machine learning. Volume 4*. New York: springer.

Blender. (2011). Blender 3d model creator. (http://www.blender.org/)

Bohg, J., Kragic, D. (2009). Grasping familiar objects using shape context. *International Conference on Advanced Robotics* (pp. 1–6).

Castrodad, A., & Sapiro, G. (2012). Sparse modeling of human actions from motion imagery. *International journal of computer vision*, *100*, 1–15.

Chan, A., Croft, E., Little, J. (2011). Constrained manipulator visual servoing (cmvs): Rapid robot programming in cluttered workspaces. *IROS* (pp. 2825–2830).

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, *43*, 129–159.

Cheng, B., Yang, J., Yan, S., Fu, Y., & Huang, T. S. (2010). Learning with l1-graph for image analysis. *IEEE Transactions on Image Processing*, *19*, 858–866.

Choi, C., Baek, S., Lee, S. (2008). Real-time 3d object pose estimation and tracking for natural landmark based visual servo. *IROS* (pp. 3983–3989).

Detry, R., Piater, J. (2011). Continuous surface-point distributions for 3d object pose estimation and recognition. *ACCV* (pp. 572–585).

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). New York: Wiley.

Eberhart, R., Shi, Y., & Kennedy, J. (2001). *Swarm intelligence. The Morgan Kaufmann Series in Evolutionary Computation*. San Francisco: Morgan Kaufmann.

Evermotion, T.M. (2012). Evermotion 3d models. (http://www.evermotion.org/)

Fergus, R., Perona, P., & Zisserman, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, *71*, 273–303.

Ferrari, V., Tuytelaars, T., & Van Gool, L. (2006). Simultaneous object recognition and segmentation from single or multiple model views. *IJCV*, *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.

Guha, T., & Ward, R. K. (2012). Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*, 1576–1588.

Hebert, P., Hudson, N., Ma, J., Howard, T., Fuchs, T., Bajracharya, M., Burdick, J. (2012). Combined shape, appearance and silhouette

for simultaneous manipulator and object tracking. In *Robotics and Automation (ICRA)*, 2012 IEEE International Conference on, IEEE (pp. 2405–2412).

Hinterstoisser, S., Benhimane, S., Navab, N. (2007). N3m: Natural 3d markers for real-time object detection and pose estimation. *ICCV* (pp. 1–7).

Hsiao, E., Collet, A., Hebert, M. (2010). Making specific features less discriminative to improve point-based 3d object recognition. *CVPR* (pp. 2653–2660).

Jolliffe, I. (1986). *Principal Component Analysis*. New York: Springer Verlag.

Kouskouridas, R., Gasteratos, A., & Badekas, E. (2012). Evaluation of two-part algorithms for objects' depth estimation. *Computer Vision, IET*, *6*, 70–78.

Kouskouridas, R., Gasteratos, A., & Emmanouilidis, C. (2013). Efficient representation and feature extraction for neural network-based 3d object pose estimation. *Neurocomputing*, *120*, 90–100.

Kragic, D., Björkman, M., Christensen, H., & Eklundh, J. (2005). Vision for robotic object manipulation in domestic settings. *RAS*, *52*, 85–100.

Krainin, M., Henry, P., Ren, X., & Fox, D. (2011). Manipulator and object tracking for in-hand 3d object modeling. *IJRR*, *30*, 1311–1327.

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

Lippiello, V., Siciliano, B., & Villani, L. (2007). Position-based visual servoing in industrial multirobot cells using a hybrid camera configuration. *IEEE Transactions on Robotics*, *23*, 73–86.

Lippiello, V., Ruggiero, F., & Siciliano, B. (2011). Floating visual grasp of unknown objects using an elastic reconstruction surface. *IJRR*, *70*, 329–344.

Lowe, D. (1999). Object recognition from local scale-invariant features. *ICCV*, *2*, 1150–1157.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, *60*, 91–110.

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

Mason, M., Rodriguez, A., & Srinivasa, S. (2012). Autonomous manipulation with a general-purpose simple hand. *IJRR*, *31*, 688–703.

Mei, L., Liu, J., Hero, A., Savarese, S. (2011). Robust object pose estimation via statistical manifold modeling. *ICCV* (pp. 967–974).

Mei, L., Sun, M., M.Carter, K., III, A.O.H., Savarese, S. (2009). Object pose classification from short video sequences. *BMVC*.

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.

Oikonomidis, I., Kyriazis, N., Argyros, A. (2011). Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. *ICCV* (pp. 2088–2095).

Pang, Y., Li, X., & Yuan, Y. (2010). Robust tensor analysis with l1-norm. *IEEE Transactions on Circuits and Systems for Video Technology*, *20*, 172–178.

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. *RAS*, *58*, 551–565.

Qiao, L., Chen, S., & Tan, X. (2010). Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, *43*, 331–341.

Rasolzadeh, B., Björkman, M., Huebner, K., & Kragic, D. (2010). An active vision system for detecting, fixating and manipulating objects in the real world. *IJRR*, *29*, 133–154.

Savarese, S., Fei-Fei, L. (2007) 3d generic object categorization, localization and pose estimation. *ICCV* (pp. 1–8).

Saxena, A., Driemeyer, J., Kearns, J., Osondu, C., Ng, A. (2008). Learning to grasp novel objects using vision. In *Experimental Robotics*, (pp. 33–42) Berlin: Springer .

Saxena, A., Driemeyer, J., Kearns, J., & Ng, A. (2006). Robotic grasping of novel objects. *Neural Information Processing Systems*, *19*, 1209–1216.

Saxena, A., Wong, L., Quigley, M., & Ng, A. Y. (2011). A vision-based system for grasping novel objects in cluttered environments. *Robotics Research*, *18*, 337–348.

Schölkopf, B., Smola, A.J., *Müller*, K.R. (1997). (pp. 583–588). Kernel principal component analysis. *ICANN*.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Cambridge: MIT Press.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *PAMI*, *22*, 888–905.

Shubina, K., & Tsotsos, J. (2010). Visual search for an object in a 3d environment using a mobile robot. *CVIU*, *114*, 535–547.

Srinivasa, S., Ferguson, D., Helfrich, C., Berenson, D., Collet, A., Diankov, R., et al. (2010). Herb: a home exploring robotic butler. *Autonomous Robots*, *28*, 5–20.

Torabi, L., & Gupta, K. (2012). An autonomous six-dof eye-in-hand system for in situ 3d object modeling. *IJRR*, *31*, 82–100.

Tsaig, Y., & Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*, 1289–1306.

Tsoli, A., Jenkins, O. (2008). Neighborhood denoising for learning high-dimensional grasping manifolds. *IROS* (pp. 3680–3685).

Viksten, F., Forssen, P., Johansson, B., Moe, A. (2009). Comparison of local image descriptors for full 6 degree-of-freedom pose estimation. *ICRA* (pp. 2779–2786).

Vonikakis, V., Kouskouridas, R., Gasteratos, A. (2013). A comparison framework for the evaluation of illumination compensation algorithms. *IST* 2013 IEEE International Conference on. (pp. 264–268).

Wang, J., Sun, X., Liu, P., She, M. F., & Kong, L. (2013). Sparse representation of local spatial-temporal features with dimensionality reduction for motion recognition. *Neurocomputing*, *100*, 134–143.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *PAMI*, *31*, 210–227.

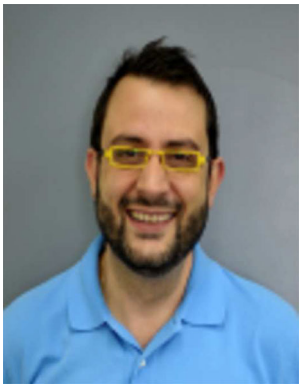
Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *PAMI*, *29*, 40–51.

Yuan, C., & Niemann, H. (2001). Neural networks for the recognition and pose estimation of 3d objects from a single 2d perspective view. *IMAVIS*, *19*, 585–592.

Zou, H., Hastie, T., & Tibshirani, R. (2004). Sparse principal component analysis. *JCGS*, *15*, 2006.

Rigas Kouskouridas holds a Diploma and a Ph.D from the Department of Production and Management Engineering at the Democritus University of Thrace, Xanthi, Greece, in 2006. He is currently a Postdoctoral research associate at the Computer Vision and Learning Lab at the Department of Electrical and Electronic Engineering of the Imperial College London. His areas of interest include computer vision, pattern recognition, machine learning and robotics. He is involved in several national (Greek) and international (European) research projects in the field of machine vision systems. Dr. Kouskouridas is a member of the IEEE, euCognition II and the Technical Chamber of Greece (TEE).

Kostantinos Charalampous holds a Bachelors (2009) and a Masters (2011) degrees both from the School of Informatics, Aristotle University of Thessaloniki. He is currently a Ph.D. candidate at the Department of Production and Management Engineering, Democritus University of Thrace.



Antonios Gasteratos is an Assistant Professor of "Mechatronics and Artificial Vision" at the DPME. He teaches the courses of "Robotics", "Automatic Control Systems", "Measurements Technology" and "Electronics". He holds a Diploma and a Ph.D. from the Department of Electrical and Computer Engineering, DUTH, Greece, 1994 and 1999, respectively. During 1999-2000 he was a Post-Doc Fellow at the Laboratory of Integrated Advanced Robotics (LIRA-Lab), DIST, University of Genoa, Italy. His research interests are mainly in mechatronics and in robot vision. He has published one textbook, 3 book chapters and more than 90 scientific papers. Dr. Gasteratos is a member of EURON, euCognition and I*PROMS European networks.