Establishing Low Dimensional Manifolds for 3D Object Pose Estimation

Rigas Kouskouridas and Antonios Gasteratos
Democritus University of Thrace
Xanthi Greece 67100
Emails: rkouskou@ieee.org gasteratos@ieee.org

Abstract—**We propose a novel solution to the problem of 3D object pose estimation problem that is based on an efficient representation and feature extraction technique. We build a part-based architecture that takes into account both appearance-based characteristics of targets along with their geometrical attributes. This bunch-based structure encompasses an image feature extraction procedure accompanied by a clustering scheme over the abstracted key-points. In a follow-up step, these clusters are considered to establish representative manifolds capable of distinguishing similar poses of different objects into the corresponding classes. We form low dimensional manifolds by incorporating sophisticated operations over the members (clusters) of the extracted part-based architecture. An accurate estimation of the pose of a target is provided by a neural network-based solution that entails a novel input-output space targeting method. The performance of our method is comparatively studied against other related works that provide solution to the 3D object pose estimation and that are based on a) manifold modeling, b) object part-based representation and c) conventional dimensionality reduction frameworks. Experimental results justify our theoretical claims and provide evidence of low generalization error when estimating the 3D pose of objects, with the best performance achieved when employing the Radial Basis Functions kernel.**

I. Introduction

The task of estimating the 3D pose of an object is among the most challenging ones in computer vision due to its practical significance and its ability to be adopted into a plethora of diverse applications. In recent years, a number of applications have primarily focused on detection and estimation of objects' pose from either a single or multiple instances for a given template [1], [2], [3]. The ultimate goal is to diffuse this technology to deliver efficient accomplishment of complex tasks, such as object manipulation, robotic navigation etc [4], [5], [6]. Despite the substantial endeavors and certain achievements made so far, no advanced computer vision system characterized with sufficient trade offs between computational burden and performance, has yet been built.

Although humans exhibit remarkable skills in estimating the relative pose of rigid objects given an initial hypothesis, such an ability is limited in contemporary computer vision systems. In this paper we attempt to address this issue, by introducing a neural network-based framework that is not only able to estimate the 3D pose of any object contained in the database, but also to generalize to unknown ones. The network is trained with numerous targets contained in several available datasets [7], [8]. Furthermore, the part-based architecture is additionally guided by the extraction of the centers of the abstracted features by applying the Neural Gas algorithm [9]. One could say that the manifold of the 3D model of an object in a known training instance is governed by the distances of the extracted clusters from one particular center as shown in Fig. I. It is palpable that the proposed method involves a new input-output mapping that reduces the dimensionality of the input vectors with good performance. A number of experimental results were executed in order, firstly, to demonstrate the performance of the present technique and, secondly, to evaluate several network architectures in known datasets.

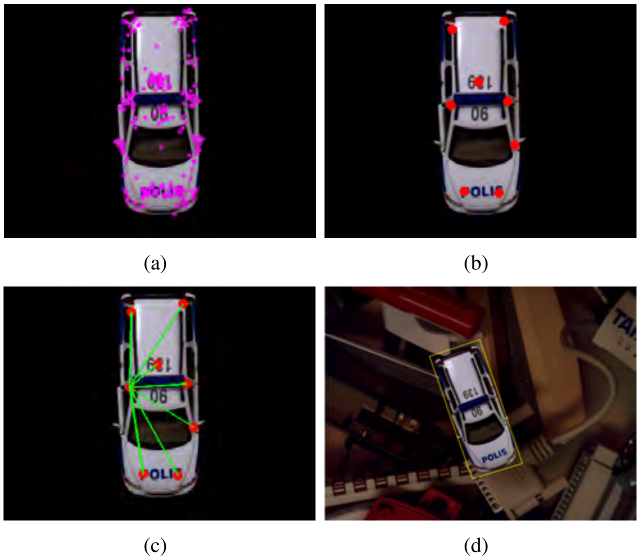

(a) (b) (c) (d)

Fig. 1. In this figure the main idea underlying the proposed method is shown. The part-based architecture is built through the processes of key-point extraction 1(a) and a clustering procedure based on Neural Gas 1(b). The next phase entails the establishment of the manifold that stands for the distance of each of these clusters with a given one 1(c). As a final step, several manifolds representing numerous training objects are considered as input to a neural network-based framework that is simulated in order to estimate the 3D pose of the object 1(d).

The contribution of this paper, among others, entails the formalization of this novel manifold modeling that avoids the use of conventional dimensionality reduction techniques widely used in image understanding applications. Additionally, the proposed method does not restrain the learning module of the algorithm by not requiring extensive supervision in

978-1-4577-1775-8/12/$26.00 ©2012 IEEE

the particular task of finding representative descriptions of a 3D object pose model. Furthermore, exhaustive experimental evaluation provided evidence of both low generalization error obtained through the proposed sparse manifold modeling and non-linear mapping between input and output space, justifying our choice to adopt a neural network-based strategy. Furthermore, we analyze the input-output mapping process and discuss generalization issues on unknown objects, while to the best of our knowledge, in the field of neural network-based computer vision, such an attribute of efficient data handling constitutes a novel approach.

II. Related work

In order to the efficiently fulfill the 3D object pose estimation task, a computer vision method should tackle several cascading issues that hinder its effective application. That have being said, the dimensionality d of the manifold of the input space that influences the performance of a regressor or a classifier is directly related with the complexity of the problem in hand. The computational burden of the training process along with costs associated with gathering, storing and processing data, can be sufficiently reduced by applying dimensionality reduction over the input vectors. Additionally, when information is represented with fewer features input models are more tolerant to noise, outliers and other disturbances, while the process is controlled in a more efficient way. Since, in this paper we adopt a neural network strategy, particular indication should be given to the dimensionality of the input vectors used for training the net. While other related approaches in this field encompass conventional dimensionality reduction techniques (e.g. Principal Component Analysis (PCA), Independent Component Analysis (ICA), Multi-Dimensional Scaling (MDS)), the proposed method is based on a new input-output mapping that targets directly the parameterized pose space. In [10] Yuan et. al, showed that a common neural network architecture accompanied by PCA for dimensionality reduction, can provide sufficient solution to the 3D object pose estimation problem. The net is trained with the resilient backpropagation method, whilst only two Degrees of Freedom (DoFs) are estimated.However, this approach fail to interpolate between two known pose configurations, since it is trained over object views with a sampling interval of 3^o and dedicated to distinguishing test patterns into the corresponding classes. Finally, we could identify several other methods [11], [12], [13] that failed to attract scholars' interest on account of their vague architecture and inadequate performance stability.

Our work imposes upon previous research endeavors conducted in the area of part-based (or constellation-based) 3D object pose estimation that emphasize in learning highly discriminative object pose models [14], [15], [16]. These methods showed that the efficiency of the 2D-3D point correspondence sub-routine is directly related to the process of selecting the most prominent visual patterns available. Recently, the linkage between input features and output score was evaluated based on a Fuzzy Functional Criterion. This fuzzy logic-based approach was presented in [17], whilst reporting a remarkable solution to the head pose estimation that depends on a large scale training module. On the other hand, the main attribute of any constellation-based scheme, as it was shown in [18], constitutes its remarkable ability to portray compact object models by linking together diagnostic "parts" of the objects from different viewpoints. These "parts" correspond to large and distinguishable regions over the surface of the objects that are composed of large amount of local invariant appearance-based features.

To the best of our knowledge and regarding the processes of feature selection and manifold modeling, the closest works to our paper are presented in [19] and [20], respectively. In [19] Hinterstoisser et. al, proposed a bunch-based structure called "Natural 3D Markers", which employs a method for extracting 4 or 5 close feature points that enjoy both distinctive photometric properties and equal distribution over the visible surface of the objects. However, as comparative experimental results prove, this method fails to construct compact and abstract representations of the 3D objects, whilst being less tolerant to partial occlusions. On the other hand, the work presented by Mei et. al in [20], provided evidence of efficient manifold modeling that enables accurate 3D object pose estimation. Notwithstanding its remarkable results, this approach has several drawbacks: a) the part-selection process requires extensive supervision during the learning procedure, whilst limiting the 3D object pose estimation to cars only; b) the resulting manifolds are of very high dimensionality that influences directly the performance of any regressor or classifier; c) the dataset used for learning is limited. Additionally, due to the fact that the learnt manifolds are of arbitrary dimensionality and architecture, the authors proposed the "alignment" and "expansion" operations, in order to credit for intra-class distance minimization and inter-pose variability maximization, respectively. In our case, we utilize a manifold modeling approach of a known architecture that does not require the aforementioned operations. Moreover, we also show that the proposed input-output mapping adopted by a neural network strategy is experimentally proved to provide more accurate results.

III. Methodology

In this section we present the input-output mapping process upon which the training of the neural network-based approach is performed. The overall system can be viewed as a mapping from a set of input variables $x = x_1, \ldots, x_d$, belonging to a feature-space $\mathcal{X}$, to a modeled output variable $y = y(x; w) \in \mathcal{Y}$, with w denoting the vector of the adjustable parameters. The ultimate goal of our system is to learn a regressor $g : \mathcal{X} \to \mathcal{Y}$ from an a priori training dataset $\{x^n, y^n\}$, in order to efficiently approximate the output $\mathcal{Y}_t$, when an unknown example $\mathcal{X}_t$ is provided. Fig. 2 illustrates the basic components of our system that are next discussed in the remainder of this section.

The labeled training dataset contains m training examples, i.e. images, of k objects-targets along with the corresponding pose groundtruth. The construction of the training set $\{x^n, y^n\}$

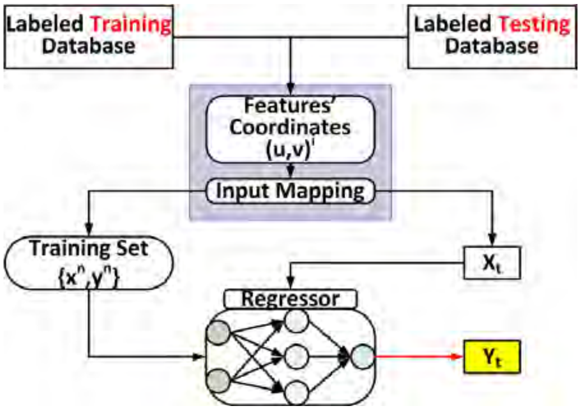


Fig. 2. Initially, labeled databases are divided into training and testing subsets, whilst for every object of the first set features' coordinates $(u, v)^i$ are extracted. As a follow-up step, the proposed input-output mapping technique takes over the construction of the set $\{x^n, y^n\}$ that is used for training the regressor. The ultimate purpose of our system is to provide an efficient approximation $\mathcal{Y}_t$ when an example $\mathcal{X}_t$, belonging to the testing subset, is presented to the network.

is based on an iterative process over m images of k objects. For the facilitation of the nomenclature and with a view to reader's better understanding, the remainder of this section presents the aforementioned iterative process for the specific object k^*. Initially, the image feature coordinates $(u, v)^i$ are calculated, with i denoting the number of the extracted keypoints. The next step is to employ the Neural Gas clustering algorithm, in order to appoint feature vector $\mathcal{E} = (u^*, v^*)^c$, where c represents the number of clusters organized as vectors and $\mathcal{E}$ the new images coordinates of the these clusters.

Let $\mathbf{e}^* \in \mathcal{E}^*$ be a randomly selected example drawn from $\mathcal{E}^* \subseteq \mathcal{E}$ vector of clusters. The proposed input-output mapping method proceeds by estimating the Euclidean distance between vector $\mathcal{E}$ and anchor point $\mathbf{e}^*$:

$$x^i = ||\mathcal{E} - \mathbf{e}^*||^2 = \sum_{i=1}^{c} \{\mathcal{E} - \mathbf{e}^*\}^2 \text{ for } i = 1, \ldots, c \qquad (1)$$

The most common approach for the input normalization is the linear transformation of given vectors so that input variables are independent. Basically, such a kind of information transformation are generally based on the mean removal method and results in sets of input vectors that have zero mean and unit standard deviation. However, this linear rescaling treats input variables as independent while in most of the cases they are not. With a view to achieve an efficient solution to this problem we adopted a more prominent strategy which allows correlations amongst variables [21]. Therefore, input variable x^i is organized into vector $\mathbf{x} = (x_1, \ldots, x_c)^{\mathbf{T}}$, while the sample mean vector and the covariance matrix with respect to the $\mathcal{L}$ data points of the training set are:

$$\overline{\mathbf{x}} = \frac{1}{\mathcal{L}} \sum_{n=1}^{\mathcal{L}} x^n$$

$$\Sigma = \frac{1}{\mathcal{L} - 1} \sum_{n=1}^{\mathcal{L}} (\mathbf{x}^n - \overline{\mathbf{x}})(\mathbf{x}^n - \overline{\mathbf{x}})^{\mathbf{T}} \qquad (2)$$

This normalization results in vectors with the input variables given by the following formula:

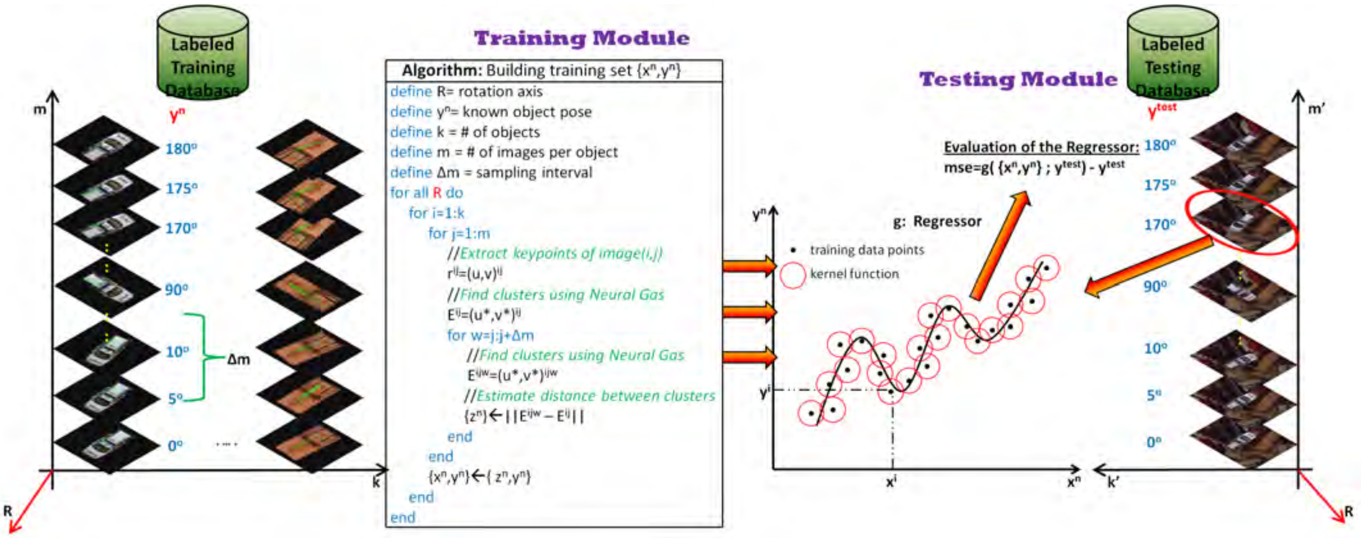
$$\widetilde{\mathbf{x}}^n = \Lambda^{-1/2}\mathbf{U}^{\mathbf{T}}(\mathbf{x}^n - \overline{\mathbf{x}}) \qquad (3)$$

where $\mathbf{U} = (u_1, \ldots, u_c)$ and $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda_c)$ correspond to the eigenvectors and eigenvalues, respectively, which are calculated from the covariance matrix $\Sigma u_\rho = \lambda_\rho u_\rho$.

The input-output mapping procedure iterates over m images of k objects holding information about the pose of the target. Since the employed databases contain numerous combinations of geometrical orientations, the most challenging task consists of finding features that repeat when matching one object's image with others depicting the same target under different viewpoints. More specifically, training datasets consist of images of objects shot every 5^o and correspond to known poses y^n , as they are placed on a turntable and the orientation of the camera alters between the three axis X, Y and $Z \in R$. In order to clarify the feature extraction process, the building of the training set and the simulation of the network we illustrate the process in Fig. 3. The training phase incorporates both the building of the training set $\{x^n, y^n\}$ and the training of the regressor. Δm as shown in Fig. 3, stands for the *tracking sensitivity* of our system and its span, e.g. $[-30^o, +30^o]$, constrains the output of the regressor to the same range, without affecting the efficiency of the tracking process though. The regressor g as shown in the particular example of Fig. 3, is a RBF-based one being responsible for estimating the pose of the test object y^{test} as $g(\{x^n, y^n\}; y^{test})$. In order to evaluate the potential of the architecture of each regressor we have examined the corresponding mean squared errors. The final stage of the proposed framework encompasses the training of the neural network-based regressor using the set $\{x^n, y^n\}$ and the simulation of its output. In order to evaluate the performance of the regressor we have tested several neural network architectures with numerous attributes.

IV. Experimental Results

Initially the authors would like to note that there is a serious lack of databases devoted to 3D object pose estimation, opposed to datasets existing for recognition and classification purposes. Furthermore, as far as the experimental evaluation is concerned, we make use of the only available databases of COIL-100 [8] and CVL [7] for 3D object pose estimation. In addition, the feature extraction process is accomplished using the SIFT algorithm [22] followed by homography-based RANSAC [23] for outliers removal. Moreover, as stated above, unsupervised clustering is accomplished through the Neural Gas algorithm presented in [9]. At this point we would like to state that, regarding the feature extraction process, the proposed framework can be easily adjusted in order to integrate any combination comprising a detector and a descriptor and it is not limited by the selection of SIFT. Likewise, concerning the clustering procedure, there are no limitations, while Neural



```
Algorithm: Building training set $\{x^n,y^n\}$
define R= rotation axis
define $y^n$= known object pose
define k = # of objects
define m = # of images per object
define Δm = sampling interval
for all R do
    for i=1:k
        for j=1:m
            //Extract keypoints of image(i,j)
            $r^{ij}=(u,v)^{ij}$
            //Find clusters using Neural Gas
            $E^{ij}=(u^*,v^*)^{ij}$
            for w=j:j+Δm
                //Find clusters using Neural Gas
                $E^{ijw}=(u^*,v^*)^{ijw}$
                //Estimate distance between clusters
                $\{z^n\} \leftarrow ||E^{ijw} - E^{ij}||$
            end
        end
        $\{x^n,y^n\} \leftarrow \{z^n,y^n\}$
    end
end
```

Fig. 3. The training module of the proposed system comprises the process of building the training set $\{x^n, y^n\}$ and that of training the neural network-based regressor. In the first instance, images of objects belonging to labeled datasets dedicated to training are drawn with the view to construct the training set to be fed to the regressor. As a final step, images of targets associated with the testing databases are further processed in order to provide an estimation of the pose of the test objects.

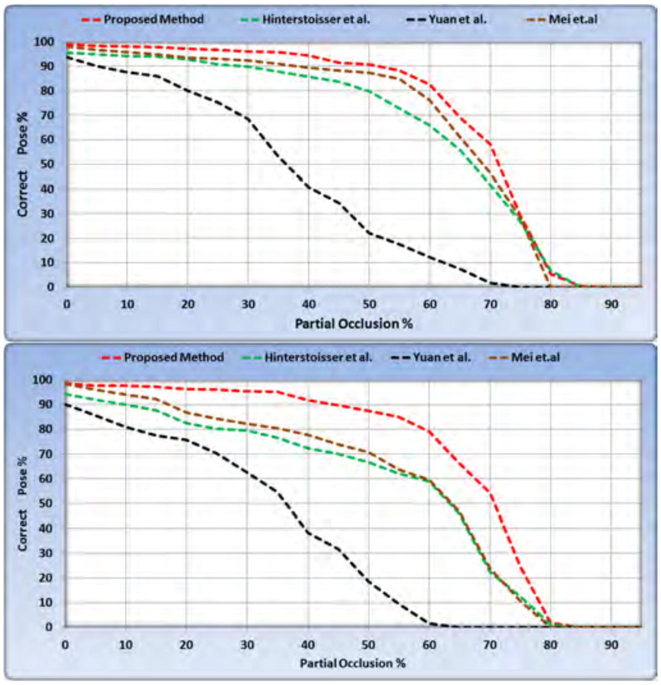


Fig. 4. Comparative evaluation of the proposed framework in cases where an estimation of the pose is considered successful if the error of the estimated rotation parameters is less than 5 (Upper) and 3 (Lower) degrees, respectively.

Gas was favored among others primarily due to its quantization capabilities.

The most common evaluation criterion utilized by methods dedicated to 3D object pose estimation is the one of testing the performance of the respective algorithms under varying percentages of partial occlusions. The latter play a vital role in image understanding applications since they affect directly the efficacy of decoding schemes. In this paper, we address this issue by expanding our training set $\{x^n, y^n\}$ with images of partially occluded targets. Partial occlusions are introduced artificially, while the percentage of obstruction lying in the range [0-95]. Moreover, by adopting the evaluation criterion presented in [19], we have tested our framework's variance against occlusions and compared it with other highly related work resulting in the outcomes depicted in Fig. 4(Upper). This figure demonstrates the condensed results from simulated networks with over 200 testing examples are presented. According to the adopted evaluation criterion, an estimation of the pose is considered successful, if the error of the estimated rotation parameters is less than 5^0. The RBF-based version of our system proved to be more tolerant to partial occlusions compared to the works of Hinterstoiser et al. [19], Mei et al. [20] and Yuan et al. [10]. Furthermore, we have additionally evaluated all methods for a permissible error of 3^o with the results being depicted in Fig. 4(Lower). The visual outcome of the proposed manifold modeling approach is presented in Fig. 7. Fig. 6 depicts the performance of our work in the particular task of tracking unregistered objects where an unknown test example was presented to the system. Finally, we have further tested our approach for totally unregistered objects in order to evaluate its performance against large oscillations in scale and illumination circumstances, with the visual results illustrated in Fig. 5.

V. Conclusion

We proposed a new manifold method that moves away from conservative dimensionality reduction schemes, such as the PCA with their inevitable information loss. Concretely, in this paper we presented a novel solution to the 3D object

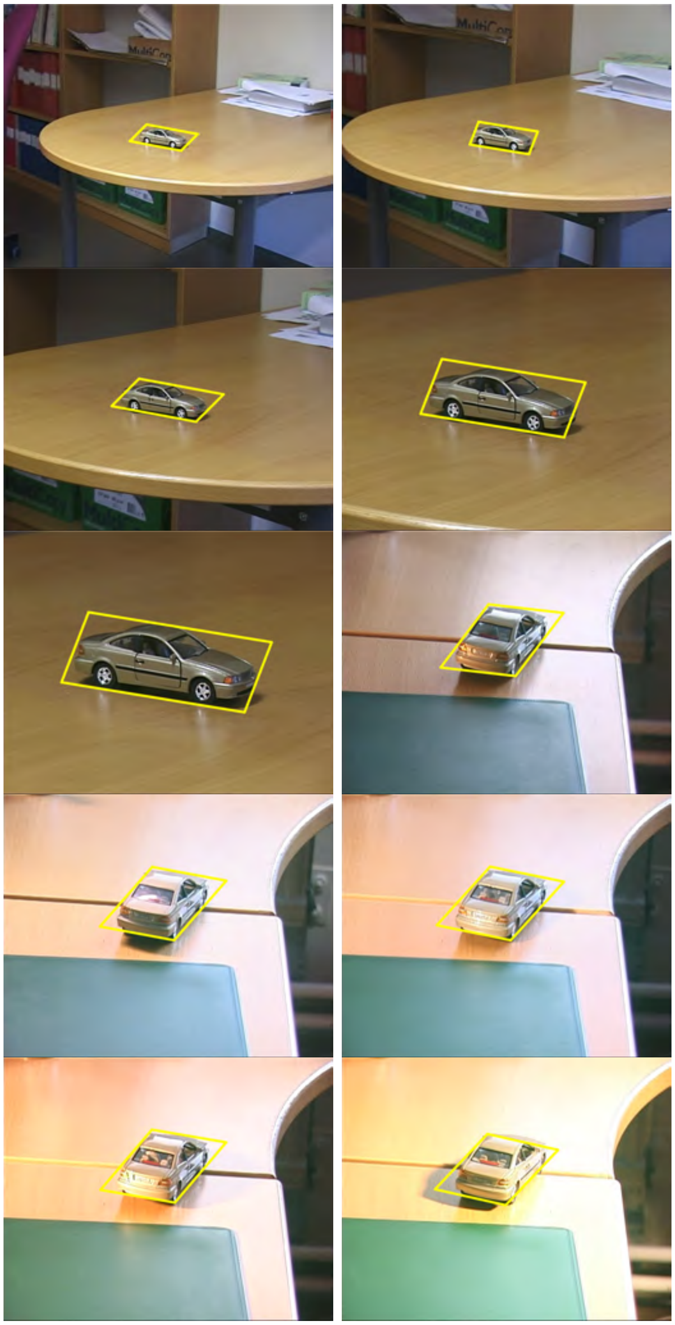


Fig. 5. The proposed approach can efficiently provide accurate estimations regarding the 3D pose of unregistered objects for different scales and illumination conditions.

pose estimation problem that lays its foundations on unsupervised learning and its part-based architecture that encapsulated both appearance and geometrical attributes of the objects. Moreover, experimental results revealed two major aspects: a) The relation between input space, which is defined over the trained objects, and the output one, characterizing the target models, is a non-linear one; b) our choice to adopt a neural network-based strategy, which efficiently captured this non-linear binding, was totally justified. Looking ahead to future work, the authors plan to construct a new database for 3D object pose estimation, with translation parameters included. In addition to, future work entails testing the performance of the system under several mixtures of detectors and descriptors and clustering techniques.

REFERENCES

[1] R. Detry and J. Piater, "Continuous surface-point distributions for 3d object pose estimation and recognition," *ACCV*, vol. 1, pp. 572–585, 2011.

[2] J. Ma, T. Chung, and J. Burdick, "A probabilistic framework for object search with 6-dof pose estimation," *The International Journal of Robotics Research*, vol. 30, no. 10, pp. 1209–1228, 2011.

[3] R. Kouskouridas, A. Gasteratos, and E. Badekas, "Evaluation of two-part algorithms for objects' depth estimation," *Computer Vision, IET*, vol. 6, no. 1, pp. 70–78, 2012.

[4] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool, "3d urban scene modeling integrating recognition and reconstruction," *International Journal of Computer Vision*, vol. 78, no. 2, pp. 121–141, 2008.

[5] B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic, "An active vision system for detecting, fixating and manipulating objects in the real world," *The International Journal of Robotics Research*, vol. 29, no. 2-3, p. 133, 2010.

[6] R. Kouskouridas, A. Amanatiadis, and A. Gasteratos, "Guiding a robotic gripper by visual feedback for object manipulation tasks," in *Mechatronics (ICM), 2011 IEEE International Conference on*. IEEE, 2011, pp. 433–438.

[7] F. Viksten, P.-E. Forssén, B. Johansson, and A. Moe, "Comparison of local image descriptors for full 6 degree-of-freedom pose estimation," *ICRA*, 2009.

[8] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia Object Image Library (COIL-100)," Columbia University, Tech. Rep., Feb 1996.

[9] B. Fritzke *et al.*, "A growing neural gas network learns topologies," *Advances in neural information processing systems*, vol. 7, pp. 625–632, 1995.

[10] C. Yuan and H. Niemann, "Neural networks for the recognition and pose estimation of 3d objects from a single 2d perspective view," *Image and Vision Computing*, vol. 19, no. 9-10, pp. 585–592, 2001.

[11] A. Saalbach, G. Heidemann, and H. Ritter, "Parametrized soms for object recognition and pose estimation," *Artificial Neural Networks*, vol. 2, pp. 140–140, 2002.

[12] A. Castro, Y. Frauel, E. Tepichín, and B. Javidi, "Pose estimation from a two-dimensional view by use of composite correlation filters and neural networks," *Applied optics*, vol. 42, no. 29, pp. 5882–5890, 2003.

[13] S. Hati and S. Sengupta, "Pose estimation in automated visual inspection using ann," *Systems, Man, and Cybernetics*, vol. 2, pp. 1732–1737, 1998.

[14] A. Berg, T. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," *CVPR*, vol. 1, pp. 26–33, 2005.

[15] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition," *International Journal of Computer Vision*, vol. 71, no. 3, pp. 273–303, 2007.

[16] D. Lowe, "Object recognition from local scale-invariant features," *ICCV*, vol. 2, pp. 1150–1157, 1999.

[17] K. Bailly and M. Milgram, "Boosting feature selection for neural network based regression," *Neural Networks*, vol. 22, no. 5-6, pp. 748–756, 2009.

[18] S. Savarese and L. Fei-Fei, "3d generic object categorization, localization and pose estimation," *ICCV*, pp. 1–8, 2007.

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

[20] L. Mei, J. Liu, A. Hero, and S. Savarese, "Robust object pose estimation via statistical manifold modeling," *ICCV*, 2011.

[21] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006.

[22] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[23] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

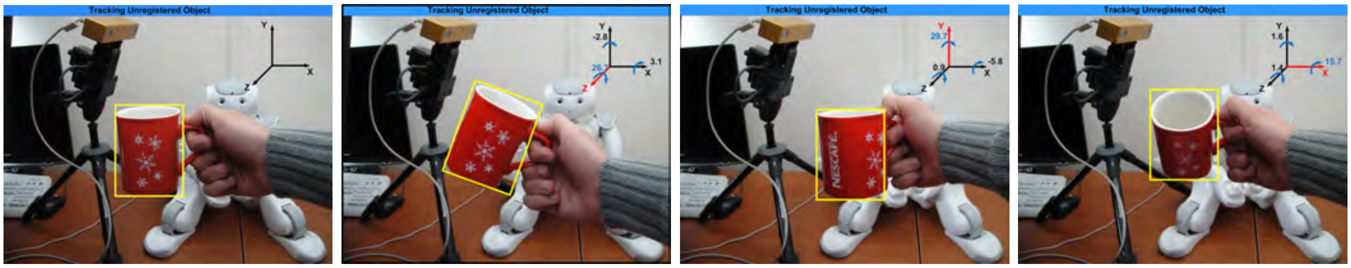

Tracking Unregistered Object
Y
X
Z
Tracking Unregistered Object
Y
-2.8
26.7
3.1
X
Z
Tracking Unregistered Object
Y
29.7
0.9
-5.8
X
Z
Tracking Unregistered Object
Y
1.6
1.4
15.7
X
Z

Fig. 6. This figure demonstrates the large generalization capacities of the proposed method since the latter is capable of tracking an unregistered object in cluttered environment under different rotations over the three axes X, Y and Z.

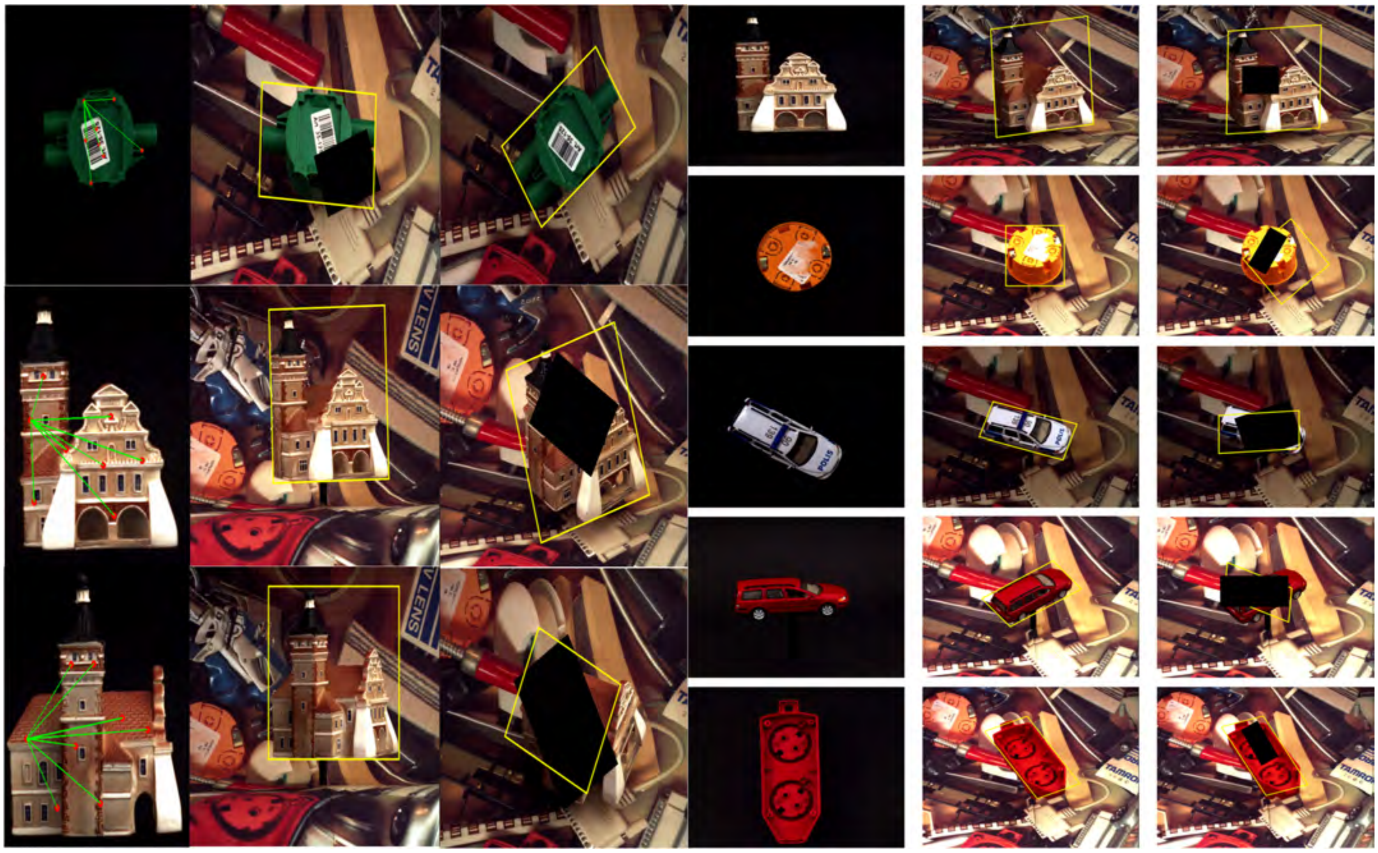


Fig. 7. This figure presents the outcome of the proposed framework for several test objects under varying percentage of artificially generated (due to database shortages) partial occlusions.