Object Recognition and Modeling Using SIFT Features

Alessandro Bruno, Luca Greco, and Marco La Cascia

DICGIM, Università degli Studi di Palermo, Italy
{alessandro.bruno15,luca.greco,marco.lacascia}@unipa.it

Abstract. In this paper we present a technique for object recognition and modelling based on local image features matching. Given a complete set of views of an object the goal of our technique is the recognition of the same object in an image of a cluttered environment containing the object and an estimate of its pose. The method is based on visual modeling of objects from a multi-view representation of the object to recognize. The first step consists of creating object model, selecting a subset of the available views using SIFT descriptors to evaluate image similarity and relevance. The selected views are then assumed as the model of the object and we show that they can effectively be used to visually represent the main aspects of the object.

Recognition is done making comparison between the image containing an object in generic position and the views selected as object models. Once an object has been recognized the pose can be estimated searching the complete set of views of the object. Experimental results are very encouraging using both a private dataset we acquired in our lab and a publicly available dataset.

Keywords: Object Recognition, Pose Estimation, Object Model, SIFT.

1 Introduction

The problem of automatically learning object models for recognition is one of the classical challenges in the field of Computer Vision. Object recognition can be formulated in terms of shape, appearance, or feature matching. In this paper we address object recognition as a features model matching problem. More particularly, we analyze the matches between local keypoints in multiple views of the same object to extract a model of the object. The SIFT [1] keypoints descriptors are used to address object recognition problem. Research in object recognition is increasingly concerned with the ability to recognize specific instances or generic classes of objects. We focus our attention on the problem of the recognition of specific instances of objects into the images.

Many methods in object recognition separate processing into two main steps: feature extraction and matching. In the first stage, discrete primitives, or features are detected. In the second stage, stored models are matched against those features.

From a neuroscientific perspective, object recognition is one of the most fascinating abilities that humans possess. It is easy (for human being) to generalize from observing a set of objects to recognizing objects that have never been seen

J. Blanc-Talon et al. (Eds.): ACIVS 2013, LNCS 8192, pp. 250–261, 2013.

before. On the contrary, it is not simple to develop vision systems that match the cognitive capabilities of human beings. The relative pose of an object to a camera, the lighting variation of a scene, and generalization from a set of exemplar images are some of the development of a vision system for object recognition should cope with. A good object recognition system should be able to extract and recognize the regularities of images, taken under different lighting and pose conditions. Object models and representations capture the most important features of the same object; furthermore the models for object recognition should be as more compact as possible to allow a lower computational complexity in the recognition phase. The representations can be either 2D or 3D. The recognition process is carried out by matching the test image against the stored object representations or models.

In this paper we present a new method for automatically learning object models for recognition. We used a dataset in which each object is subjected to rotation of 180 degrees, in steps of 5 degrees along yaw-axis, and 90 degrees along pitch-axis, in steps of 5 degrees (703 image samples for each object). We analyze the number of matches between nearby images (rotations of 5 degrees) by choosing, as a model of the object, only the most representative images (the criterion of choice of the images will be described in greater detail in section 3). As result we create a model for each object that has a compact representation (a few images instead of 703). After the model is built, we use the models for object recognition; the pose of the object is also estimated with a good level of accuracy.

The contributions of this paper are a new method for automatically learning object models, and a new method for object recognition and pose estimation. The rest of the paper is organized as follows: section 2 gives an overview of the state of the art, sections 3 describe in detail the proposed method, and the datasets used for testing and training our system, in section 4 the experimental results are shown, section 5 ends the paper with conclusions e future works.

2 State of the Art

The most important object recognition approaches can be subdivided in three main categories:

1)Geometry-based approaches [2] [3];
2)Appearance-based algorithms [4];
3)Feature-based algorithms [5] [6].

In Geometric based approaches the main idea is that the geometric description of a 3D object allows the projected shape to be accurately analyzed in a 2D image under projective projection, thereby facilitating recognition process using edge or boundary information.

The most notable appearance-based algorithm is the eigenface method [4] applied in face recognition. The underlying idea of this algorithm is to compute eigenvectors from a set of vectors where each one represents one face image as a raster scan vector of gray-scale pixel values. The central idea of feature-based object recognition

algorithms lies in finding interest points, often occurred at intensity discontinuity, that are invariant to change due to scale, illumination and affine transformation.

3D object recognition is an important area in computer vision and pattern recognition and mainly include two steps: Object Detection and Object Recognition.

Object recognition algorithms based on views or appearances are a hot research topic [7] [8]. In [9] Pontil et al. proposed a method that recognize the objects also if the objects are overlapped. In recognition system based on view, the dimensions of the extracted features may be of several hundreds. After obtaining the features of 3D object from 2D images, the 3D object recognition is reduced to a classification problem and features can be considered from the perspective of pattern recognition. In [10] the recognition problem is formulated as one of appearance matching rather than shape matching. The appearance of an object depends on its shape, reflectance properties, pose in the scene and the illumination conditions. Shape and reflectance are intrinsic properties of the object, on the contrary pose and illumination vary from scene to scene. In [10] the authors developed a continuous and compact representation of object appearance that is parameterized by object pose and illumination (parametric eigenspace, constructed by computing the most prominent eigenvectors of the set) and the object is represented as a manifold. The exact position of the projection on the manifold determines the object's pose in the image. The authors suppose that the objects in the image are not occluded by others objects and therefore can be segmented from the remaining scene.

In [11] the author developed an object recognition system based on SIFT descriptors [1]. The features of SIFT descriptors are invariant to image scaling, translation and rotation, partially invariant to illumination changes and affine or 3D projection. SIFT are efficiently detected through a filtering approach that extract stable points in scale space. The SIFT keypoints are used as input to a nearest-neighbor indexing method, this identifies candidate object matches.

In [12] the authors analyzed the features which characterize the difference of similar views to recognize 3D objects. Principal Component Analysis (PCA) and Kernel PCA (KPCA) are used to extract features and then classify the 3d objects with Support Vector Machine (SVM). KPCA-SVM, PCA-SVM performances on Columbia Object Image Library (COIL-100) have been compared. The best performance is achieved by SVM with KPCA. KPCA is used for feature extraction in view-based 3D object recognition. In [12] different algorithms are shown by comparing the performances only for four angles of rotation (10° 20° 45° 90°). Furthermore, the experimental results are based only on images with dimensions 128 x 128.

Peng Chang et al. [13] used the color co-occurrence histogram (that adds geometric information to the usual color histogram) for recognizing objects in images. The authors computed model of Color Co-occurrence Histogram based on images of known objects taken from different points of view. The models are then matched to sub-regions in test images to find the object. Moreover thy developed a mathematical probabilistic model for adjusting the number of colors in Color Co-occurrence Histogram.

Many object recognition methods perform also object pose estimation. In [14] Kouskorida et al. proposed a solution to the problem of 3D object pose estimation. More particularly, the authors build an architecture based on appearance and geometrical attributes. The feature extraction procedure is accompained by a clustering scheme over the key-points. The clusters are considered to establish representative manifolds. In [15] Viksten et al. performed comparison of local image descriptors for full 6 Degree-of-Freedom Pose Estimation. In [16] Pose Estimation is treated as a regression problem. In [17] the authors addressed the challenging problem of pose recognition using simultaneous color and depth information. They used a multi-kernel approach to incorporate depth information to perform more effective pose recognition on table-top objects.

Our method is a new object recognition algorithm based on visual object models, it also performs pose estimation of the object. In [11] SIFT keypoints and descriptors are used as input to a nearest-neighbor indexing method that identifies candidate object matches, we, instead, used SIFT for obtaining the object model for multiple views and multiple images of the same object. In our method the recognition of the object is performed by matching the keypoints of the query image only with the keypoints of the objects models. Similarly to the Peng Chang et al. method [13] we used object modeling for object recognition but we preferred to extract local features (SIFT) rather than global features such as the color Co-occurrence histogram.

Once recognition phase is done, we estimate the pose of the object by matching the SIFT of the query image with the SIFT keypoints of all the views of that object. On the other side, Kouskurida et al. [14] proposed a method for pose estimation in which appearance and geometrical attributes are extracted and clustered over keypoints.

3 Object Modeling and Recognition

The proposed method is a recognition algorithm based on visual object models. Object models are pre-calculated starting from a particular type of image dataset.

Fig. 1. Example of a car for -90, 0 and +90 degrees rotation on a turntable

3.1 Objects Image Dataset

The method works with a particular type of dataset, that is a collection of multi angle views of each objects. For each object, the dataset contains N views from a fixed camera generated rotating the object by a fixed angle, having only one degree of

freedom (i.e. using a turntable). The same result can be obtained by taking image rotating uniformly around the fixed object. An example of this is show in Fig. 1.

If the N images cover only a specific range of views of the object (i.e. 0-180 degrees), the recognition algorithm is reliable on this range.

For each object, the model is obtained as follows:

1. SIFT keypoints and descriptors are calculated for every view;
2. For each view, only the union of the subsets of the keypoints that match with the previous and the next view is used as view descriptor;
3. Starting from image 1 to N, the number of matching keypoints of the selected subset and the subset of next view is calculated and associated to the current image;
4. The views corresponding to local minima and maxima of this sequence are selected as model of the object.

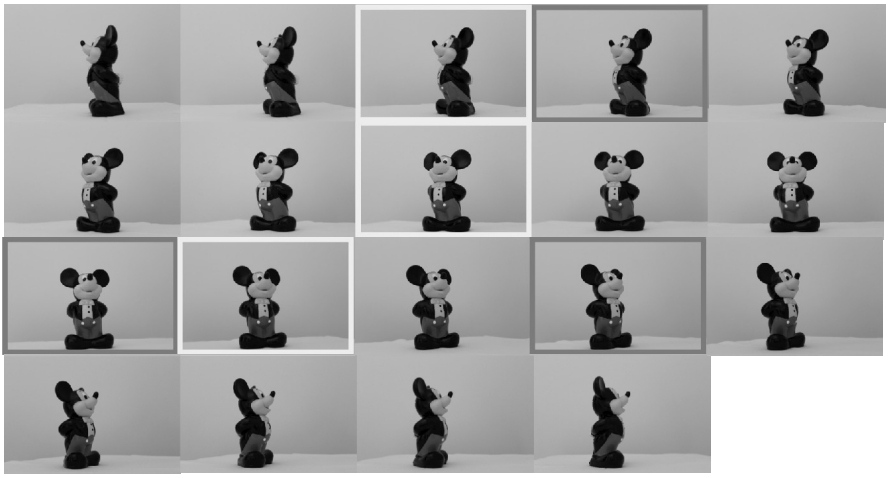

Fig. 2. Complete object view in our dataset. Squares are the subset of the image selected for the object model.

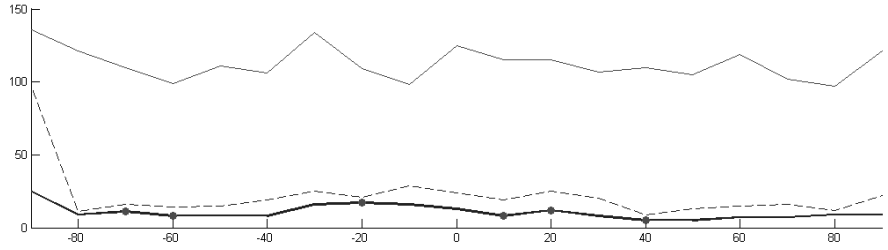

Fig. 3. The higher line is the number of detected SIFT keypoints, the middle line is the number of filtered keypoints, the lower line is the number of match between a view and the next. Circles represents maxima and minima included in the model.

Step 2 performs a filtering on keypoints keeping only those that are present in a filtering sliding window of 3 images. So only repeated and not occluded points are present in the resulting subset. In step 3 the visual continuity of the sequence is evaluated for each image looking at the match between the image and next view, so calculating the similarity in SIFT descriptors.

Taking local minima of this similarity, the corresponding images are the most dissimilar in their neighborhood, so representing views that contains a visual change of the object. Local maxima, conversely, correspond to images that contains common feature in their neighborhood, so being representative of this.

The views corresponding to local maxima and minima, so taking the images that contain "typical" views (maxima) and visual breaking views (minima). The number M of this image is lower than original N dimension of starting dataset. The value of M depends on the shape of the object.

Although the visual model for recognition is composed of the selected images, the filtered subset of descriptors is also stored to describe completely the objects and to perform pose estimation.

3.2 Dataset

To evaluate the performance of modeling objects, a dataset has been created. It contains 18 objects and 19 views for each of this, in the range [-90 90] degrees with a change of 10 degree at each step. An example of this dataset is shown in Fig. 2. The squares represent the images forming the model: the brighter ones are the one associated to local maxima, the darker ones those associated to minima.

The 18 objects are different in shape and color and so cover a very large number of possible typology of recognition scenarios.

Using this dataset the overall amount of images of the models is 93, starting from a full dataset dimension of 342.

3.3 Recognition

The recognition algorithm is based on the models of objects obtained with the procedure of the previous paragraph.

Having a new query image, containing (or not) an object present in the dataset, the recognition algorithm is:

1. Calculate the SIFT keypoints and descriptors for the query image;
2. Match the keypoints with all the filtered keypoints of the images of all models;
3. Select the object referring to the best match (over a fixed threshold, 15 in our experiments) as the recognized object.

Using this method, the number of the match to calculate is reduced to the dimension of model dataset. In the case of our dataset, the reduction is from 342 to 93, so having only the 27% of matches to calculate compared to the full dataset. Object models on average are made of 5.1 images.

3.4 Pose Estimation

Having a well-defined type of dataset, where the pose of the object is known, it is possible to have a pose estimation of the recognized object without taking explicitly into account the shape characteristics of the query image (i.e. segmentation and 3d structure of the object). In fact, once the object is recognized this can be compared with the original full model of N images of the dataset (without model filtering) to recover the best match and so the pose.

In summary, given a query image, the process of recognition and pose estimation is the following:

1. Calculate the SIFT of query image;
2. Compare with models and recognize if there is a known object;
3. If recognized, compare the query image with all images (N) of the original dataset for the recognized object and determine the pose with the largest number of matches.

The overall process consist of a number of comparison that is the sum of the models images and N. In our dataset this number is 112, so only the 33% of the comparison using the entire dataset.

Fig. 4. Result of recognition and pose estimation. Size and light condition are not the same than in dataset.

4 Results

Datasets are a key factor in recognition task when the method doesn't use external knowledge on objects (i.e. 3D information on shape and geometry). Ponce et al. in [19] shows that current datasets suffer some limitation in the number of objects available and in objects views variability. To avoid this limitation the presented method uses only a well-defined type of dataset. To analyze the recognition performance of the proposed method the algorithm has been tested on using the

dataset present in [18]. This is composed of 16 object with views with 2 degrees of freedom: the object is rotated using a turntable with shifts of 5 degree from 0 to 180 and the camera is rotated from frontal view (0 degrees) to upper view (90 degrees) with shifts of 5 degree. So there are 37 images from 19 point of views, obtaining 703 images for each object.

In our test we used only the images from the middle (45 degrees) vertical camera position and all the horizontal camera position. The overall number of image of the extracted dataset is 592.

In [18], the same images are taken with a black background and a cluttered background. In this case the testset is composed by the cluttered images from the same or from shifted camera position. In Fig. 5 are shown images from dataset and testset.

Fig. 5. From left to right: House (0°), House (30°), Cluttered House (30°), Cluttered House (30°) and camera at 30° (15° shift than dataset)

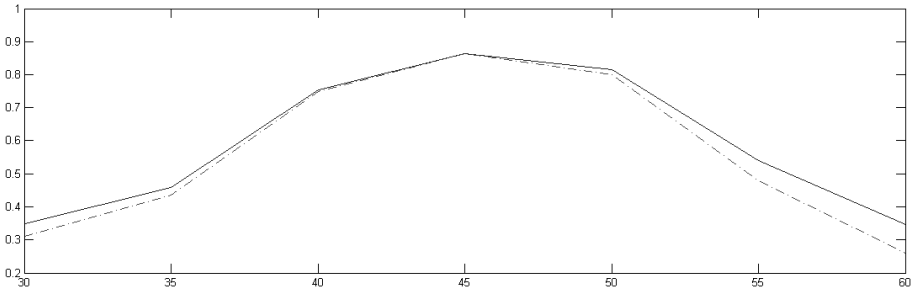

Fig. 6. Accuracy of recognition changing testsets. Solid line is recognition accuracy, dashed line is pose estimation accuracy. Best results are for testset whit the same camera angle of the original dataset. Accuracy decreases if test image are largely shifted from this angle.

4.1 Recognition and Pose Estimation Performances

The algorithm has been tested with our dataset and the reduced version of [18]. In the first case, testset is composed by query images containing objects of the dataset. Using this, the recognition performance is 80% and the pose estimation performance is 75%.

In the second case, testset is composed by cluttered images taken from vertical camera position in range [30 60], so having a displacement of 15 degrees from the fixed dataset point of view of 45 degrees. Recognition and pose estimation accuracy for central point of view is both 86%. Fig. 6 shows how performances change when the camera position of test image change respect to the original dataset position.

The same experiment was repeated with the same testset but with a random resize of the images with a proportion from 0.2 to 1. In Fig. 7 there is the plot of the results

in this case. Accuracy is uniformly lower than in the first case (73% for recognition and 72% for pose estimation) but the trend of results is very similar.

Modeling the [18] dataset, the original (592) number of images for the recognition is reduced to 217 images, with an average value of 13,5 images for object. So, the recognition task is performed with only the 36% of the total comparison. The pose estimation step adds 37 match, so the total reduction of the complete elaboration is to 43%. Table 1 and 2 report the performance for the used datasets.

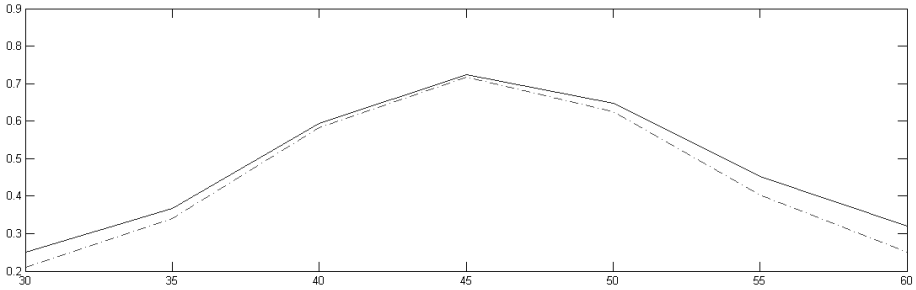

Fig. 7. Accuracy for randomly resized images of testset

Table 1. Performances of the proposed method

dataset	Recognition accuracy	Pose estimation accuracy
Our dataset	80%	75%
[18]	86%	86%
[18] random resize	73%	72%

Table 2. Comparison reduction of the proposed method

dataset	Images/model	reduction
Our Dataset	5,1	27%
[18]	13,5	37%

4.2 Limits of Recognition

Recognition by SIFT descriptors works reliably if object actually have recognizable features like texture, corners or writings. Results reported for the dataset [18] are calculated using the entire dataset, but not all the objects are really suited for recognition with proposed method.

As shown in Fig. 8 accuracy of recognition for each object is very close to one if it has shape characteristics recognizable by SIFT keypoints. Only a few objects show poor accuracy and actually they have not sufficient visual features to be recognized

using SIFT. Removing these objects from the dataset the proposed recognition method performance increases. For example, using [18] testset without the two objects shown in Fig.8, recognition accuracy increases to 96% for the same camera point of view of dataset.

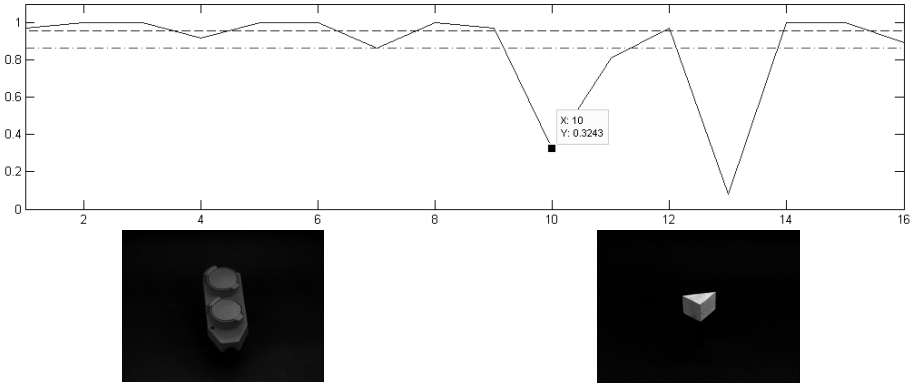

Fig. 8. Performance results versus object id number. Accuracy is low only with two objects (id 10 and 13, shown from left to right under the chart). Straights lines show average accuracy without (the higher) and with (the lower) these objects.

5 Conclusions and Future Works

We have presented an object recognition technique based on object modeling using local features. For this task, we used a multiview dataset of objects. The method performs modeling selecting only a subset of the views, creating in this way a compact representation. View selection is done by analyzing views similarity for each object by comparing SIFT descriptors.

Recognition is done by comparing a query image to the extracted models, reducing the number of comparisons with respect to the full dataset. Only pose estimation step needs to be performed using all views of the object.

Our method shows good performance in terms of accuracy for both object recognition and pose estimation. The construction of the model of objects with different shape and appearance was performed using the SIFT descriptors which extract informations only when the objects show texture surfaces, contours, edges, local maxima and minima of intensity. The worst results in term of precision for object recognition correspond to those objects that do not show regions with texture. As consequence, these objects include a few SIFT points and then it is not possible to construct a valid model for the object. In future work it would be appropriate to use descriptors of features that are present in objects that do not show texture (e.g. color descriptors, histogram, etc.) in the construction of the object model. The approach used in this paper may be easily extended to the recognition and modeling of objects in video or by using datasets with multiple degrees of freedom. Furthermore, it would be interesting to extend our work to a setting with different local keypoint descriptors

such as ASIFT, MSER, Harris Affine, Hessian Affine and SURF, described in [20] [21]. A performance evaluation of different local descriptors could give important informations, in order, to know which of them works better. These possible extensions are currently under development in our lab.

Acknowledgement. This paper has been partially supported under the research program P.O.N. RICERCA E COMPETITIVITA' 2007-2013, project title SINTESYS - Security INTElligence SYStem, project code PON 01_01687.

References

1. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
2. Mundy, J., Zisserman, A.: Geometric invariance in computer vision. MIT Press, Cambridge (1992)
3. Mundy, J.L.: Object recognition in the geometric era: A retrospective. In: Ponce, J., Hebert, M., Schmid, C., Zisserman, A. (eds.) Toward Category-Level Object Recognition. LNCS, vol. 4170, pp. 3–28. Springer, Heidelberg (2006)
4. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
5. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of IEEE International Conference on Computer Vision, pp. 1470–1477 (2003)
6. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)
7. Zhao, L.W., Luo, S.W., Liao, L.Z.: 3D object recognition and pose estimation using kernel pca. In: Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, China, pp. 3258–3262 (2004)
8. Wang, X.Z., Zang, S.F., Li, J., et al.: View-based 3d object recognition using wavelet multi-scale singular value decomposition and support vector machine. In: Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition, Beijing, pp. 1428–1432 (2007)
9. Pontil, M., Verri, A.: Support vector machines for 3D object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(6), 637–646 (1998)
10. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-D objects from appearance. International Journal of Computer Vision 14(1), 5–24 (1995)
11. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. (2), pp. 1150–1157 (1999)
12. Wu, Y.J., Wang, X.M., Shang, F.H.: Study on 3D Object Recognition Based on KPCA-SVM. In: International Conference on Information and Intelligent Computing IPCSIT, vol. 18, pp. 55–60 (2011)
13. Chang, P., Krumm, J.: Object recognition with color cooccurrence histograms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1999)
14. Kouskouridas, R., Gasteratos, A.: Establishing low dimensional manifolds for 3D object pose estimation. In: IEEE International Conference on Imaging Systems and Techniques (IST), pp. 425–430 (2012)

15. Viksten, F., Forssén, P.-E., Johansson, B., Moe, A.: Comparison of local image descriptors for full 6 degree-of-freedom pose estimation. In: IEEE International Conference on Robotics and Automation, ICRA 2009, pp. 2779–2786 (2009)
16. Torki, M., Elgammal: A Regression from local features for viewpoint and pose estimation. In: IEEE International Conference on Computer Vision (ICCV), pp. 2603–2610 (2011)
17. El-Gaaly, T., Torki, M.: RGBD object pose recognition using local-global multi-kernel regression. In: IEEE 21st International Conference on Pattern Recognition, pp. 2468–2471 (2012)
18. http://www.isy.liu.se/cvl/research/objrec/posedb/datasets.html
19. Ponce, J., Berg, T.L., Everingham, M., Forsyth, D.A., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B.C., Torralba, A., et al.: Dataset issues in object recognition. Toward Category-Level Object Recognition, 29–48 (2006)
20. Morel, J.-M., Yu, G.: ASIFT: A new framework for fully affine invariant image comparison. SIAM Journal on Imaging Sciences 2(2), 438–469 (2009)
21. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Computer Vision and Image Understanding 110(3), 346–359 (2008)