Simultaneous Visual Object Recognition and Position Estimation Using SIFT

Rigas Kouskouridas, Efthimios Badekas, and Antonios Gasteratos

Democritus University of Thrace,
Department of Production and Management Engineering,
Vas. Sofias 12, 67100 Xanthi, Greece
{rkouskou,agaster}@pme.duth.gr, badekas@anadelta.com

Abstract. In the last decade, pattern recognition tasks have flourished and become one of the most popular tasks in computer vision. A wealth of research focused on building vision systems capable of recognizing objects in cluttered environments. Moreover industries address all their efforts to developing new frameworks for assisting people in everyday life. The need of robots working closely to human beings in domestic workplaces, makes a necessity the usage of intelligent sensorial systems that are able to find patterns and provide their location in the working space. In this paper a novel method able to recognize objects in a scene and provide their spatial information is presented. Furthermore, we investigate how SIFT could expand for the purposes of location assignment of an object in a scene.

Keywords: Object Recognition, Position Estimation, SIFT, depth estimation, robot vision.

1 Introduction

Recognizing objects in a scene is one of the oldest tasks in computer vision field and still remains one of the most challenging. Every pattern recognition technique is directly related with the decryption of information contained in the natural environment. During the last decade, remarkable efforts were made to build new vision systems capable of recognizing objects in cluttered environments. Moreover, emphasis was given to recognition systems based on appearance features with local estate [1], [2]. Insensitivity against rotation, illumination and viewpoint changes, constitute the attributes of algorithms extracting features with local interest. Nowadays, vision systems are equipped with such attributes and, as a result, more and more techniques are included in industrial products.

In the last few years, a tendency to introduce autonomous robots into domestic environments is discerned. Industries address all their efforts to developing machines capable of assisting people to everyday life. To this end, a compulsory requirement is that robots must avoid obstacles in both static and dynamic environments. As a result, the sense of depth constitutes the essential attribute of such frameworks. In turn, over the past decade, significant research efforts were

M. Xie et al. (Eds.): ICIRA 2009, LNAI 5928, pp. 866–875, 2009.
© Springer-Verlag Berlin Heidelberg 2009

devoted to the development of vision systems, capable of providing a sense of location and direction to robots. In this field, techniques for visual Simultaneous Localization and Mapping (vSLAM) [3] [4], are included. Particularly, SLAM methods that are utilized in autonomous robotics, concentrate at building up maps, in unknown environments, whilst keeping track of their own position.

Furthermore, from time to time, researchers emphasized in introducing computer vision techniques into demanding robotics applications. As a result, challenging automatic manipulation tasks can be adequately accomplished by utilizing object recognition techniques based on local appearance. The most favored method of this field is the Scale Invariant Feature Transform (SIFT) that was presented in [5]. Furthermore, SIFT is adopted in modern robotics applications due to the fact that it performs exceptional repeatability and invariance against possible illumination, scale, rotation and viewpoint changes. For instance in [6], a remote-interactive mode for a museum guide robot is presented, where SIFT is used during the object recognition process. In turn, mobile robots' navigation using landmarks can be adequately fulfilled by using SIFT features as it was shown in [7], where each landmark is initially located in the image coordinates by using SIFT features for the recognition and the RANSAC(Random Sample Consensus) for the matching procedure.

In this paper, we describe a novel method to develop a simultaneous visual object recognition and position estimation system at the same time. We investigate how, a very efficient object recognition scheme, can be expanded for the needs of position estimation. Specifically, SIFT was selected among a set of high-level algorithms to describe patterns and objects. We prove that information derived from SIFT, allows the estimation of the distance between camera and objects found in a scene and we describe the respective algorithm. The remainder of this paper is structured as follows: In Section 2, a short introduction to object recognition based on local features is presented and a detailed description of the methods involved in building our recognition and position estimation algorithm, is given. In Section 2.1 we present the parts of SIFT that are used in the later stages of the algorithm. The proposed object recognition and position estimation framework is analytically presented in Section 3, where details about training and implementation of our system are apposed. In Section 4, we experimentally evaluate the proposed method. The work concludes with some final notes and an outlook to future work in Section 5.

2 Local Appearance-Based Recognition

During the past decade, several techniques that enforce the essential role of local features in object recognition tasks [8] were presented. The special visual distinctiveness of an object in a scene is ensured by locally sampled descriptions. The vital issue underlying object recognition based on local features is maintaining this distinctive regional-based information. Detectors and descriptors of areas of interest constitute the sub-mechanisms in every local-based recognition approach. In addition, they provide special attributes such as, insensitivity

against rotation, illumination and viewpoint changes. In [9], local detectors and descriptors are evaluated for object recognition purposes.

The main idea behind interest location detectors is the pursuit of points or regions with unique information in a scene. These spots or areas contain data that distinguish them from others in their local neighborhood. Needless to say that, detector's efficiency relies on its ability to locate, as many distinguishable areas as possible, in an iterative process. One of the most efficient detectors, the Maximal Stable Extremal Regions Detector, was proposed in [10]. In short, regions darker or brighter than their surroundings are detected. The efficiency of the algorithm relies on the relationship between pixels' intensity value and local neighborhood.

Generally speaking, a descriptor organizes the information collected from the detector in a discriminating manner. Thus, locally sampled feature descriptions are transformed into high dimensional feature vectors. In other words, parts of an object located in a scene are represented by descriptors. Putting these descriptors in logical coherence fulfills the final object representation. In all the recent proposed methods, databases containing descriptors from multiple objects are constructed. These databases, that play essential role in all the recently proposed object recognition techniques, are structured in a vocabulary-tree format. Furthermore, vocabulary trees, which are tree-like data structures based on k-means clustering, were proposed in [1].

Speeded Up Robust Features (SURF) that was introduced in [11], implements both a detector and a descriptor. The first is constructed by using a so-called Fast Hessian Matrix that, is based on an approximation of the Hessian Matrix for a given image point. Afterwards, rectangular -9x9-pixels filters are used for the approximation of the second derivative of the Gauss function. The descriptor is produced based on the responses of all the interest points extracted from the detector. Currently, the most widely adopted approach that produces efficient detector and descriptor, is SIFT [5]. As it is mentioned before, the proposed method is based on this approach. For this reason an extended description of SIFT follows.

2.1 Description of SIFT

Initially, the image is convolved with the variable-scale Gaussian for the production of a scale-space image. Afterwards, stable keypoint locations are detected by using scale-space extrema in the difference-of-Gaussian (DoG) function convolved with the image. In [12] it was shown that, DoG function provides a close approximation to the scale-normalized Laplacian of Gaussian. Thus, the scale invariance of the detector is ensured. SIFT's descriptor is produced by using stacked gradient histograms over 4x4 sample regions. Firstly, the gradient magnitude and orientation at each point in a region around the keypoint location are computed. Afterwards, these samples are gathered into orientation histograms collecting the contents over 4x4 sub-regions. Since 8 orientation bins are used, the descriptor, finally, constitutes of 128-element feature vector. In the most the cases, matching between descriptors relies on comparing them one by one. The

matching process in SIFT involves the organization of descriptors from trained images into a vocabulary kd-tree. Moreover, with this way the approximate nearest neighbors to the descriptors are found. SIFT's detector and descriptor are insensitive to possible image scale, rotation, change in 3D viewpoint, addition of noise and change in illumination. These exceptional attributes justify the fact that SIFT is adopted from, almost any, new object recognition approach [13] [14]. To sum up, SIFT's unique properties can be used not only for recognition but also for position estimation of an object in a scene. We are examine how SIFT could be altered for the purposes of such a challenging task.

3 Algorithm Description

Initially, we present fragmentarily the proposed method by showing its main stages. The main idea behind the proposed method is to maintain SIFT's properties whilst exploiting them in order to particularly estimate objects' distance from the camera. Thus, we have constructed a large database containing images from several objects. With a view to database's enrichment, these objects were photographed from different viewpoints and distances from the camera. Moreover, we used SIFT's matching sub-procedure to build an on-line scene search engine. Estimations derived from this engine are taken into account for the position estimation task. The main stages of the proposed algorithm are as follows:

Stage I *Apply SIFT to the scene's and object's image, in order to estimate the features position in each of them.*

Stage II *Obtain the N features that match in the two images by applying the matching sub-procedure of SIFT. Define as* $(X_{S_i}, Y_{S_i}), i = 1, \ldots, N$ *the positions of the N features in the scene image and* $(X_{0_i}, Y_{0_i}), i = 1, \ldots, N$ *the positions of the N features in the object image.*

Stage III *Define as* (X_{S_c}, Y_{S_c}) *and* (X_{0_c}, Y_{0_c}) *the features' centers of mass for both images. This is accomplished by estimating the mean values of the features positions in the two images:*

$$X_{S_c} = \frac{1}{N}\sum_{i=1}^{N} X_{S_i} \quad \textit{and} \quad Y_{S_c} = \frac{1}{N}\sum_{i=1}^{N} Y_{S_i}$$

$$X_{0_c} = \frac{1}{N}\sum_{i=1}^{N} X_{0_i} \quad \textit{and} \quad Y_{0_c} = \frac{1}{N}\sum_{i=1}^{N} Y_{0_i}$$

Stage IV *Calculate the mean Euclidian distance (in pixels) of each feature from the corresponding center of mass that is extracted in the previous stage. Set as* E_S *and* E_0 *the*

mean Euclidian distances in the scene and object image, respectively. The following relations are used:

$$E_S = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(X_{S_i} - X_{S_c})^2 + (Y_{S_i} - Y_{S_c})^2}$$

$$E_0 = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(X_{0_i} - X_{0_c})^2 + (Y_{0_i} - Y_{0_c})^2}$$

Stage V *Estimate d_S which corresponds to the ratio of the two mean distances E_S and E_O. Furthermore, we introduce the pre-computed depth d_O, which is obtained, during the training session and while the object is captured alone.*

$$d_S = \frac{E_0}{E_S}$$

Stage 1 could be apprehended as the training session of our algorithm. In this phase, for each image in the database keypoint features are extracted using SIFT. Each object is photographed at different distances from the camera and the pre-computed depth d_O is stored for further exploitation. This process is performed while the system is offline, thus, executable time is not taken into account. The results are stored for further use at the next phases. In Stage 2, the matching sub-procedure of SIFT is performed. Especially, descriptors that are common in both images (scene and object) are extracted. It is apparent that, one image representing the scene is compared with several others, representing the object from different viewpoints. Furthermore, the locations of the common features are stored for further exploitation.

In Stage 3, the position estimation sub-procedure takes place till the end of the algorithm. Moreover, at this phase, the features' centers of mass in both images are calculated. The last is obtained by estimating the mean values of features locations in both representations. In Stage 4, the distance of each keypoint from the center of mass is calculated. This is measured in pixels with the use of Euclidian Distance. By the end of this sub-routine, we are able to collect significant spatial information of an object in a scene. This is accomplished by simply estimating the distribution of trained features around their center of mass. Finally, in Stage 5, the object's distance from the camera is computed. The pre-computed depth d_O measured during the training session (Stage 1), is taken into account. The ratio d_S is used to measure the proportion of object's features to those found in the scene.

After the necessary training session and the database construction at the initial stages of the method, an on-line search engine follows. This is responsible for querying in the scene for objects contained in the trained database. When an object is found, the scene's image is compared to this object, which provides the majority of common matches. Finally, features' information from both images is interpolated with a view to object's position allocation.

4 Experimental Results

In this section, we assess the properties of the proposed method in detail. The tests were executed on a typical PC with a core2duo 2.2 GHz processor, 2 GB RAM and Windows XP operating system. Furthermore, the camera used (Grasshopper by Point Grey Research) is able to capture images up to 1280x960 pixels resolution and is connected to the PC via a firewire port. The data transmission is accomplished by using IEEE 1394b transfer protocol. The training of the object recognition system is done in MATLAB. The last is preferred from other programming tools, due to the fact that, it offers users-friendly environment and convenient image processing functions.

The proposed method is evaluated through exhaustive tests containing several scenes and objects. In Figure 1(a) we present scene A, which contains three different objects (e.g. a book, a modem's box and a motherboard's box). With a view to reader's better understanding, objects found in the scene are referred as book, modem and box, respectively.

(a) (b)

Fig. 1. (a) Upper: Scene A containing three objects Lower: Each object is captured from different viewpoint. (b) The extracted SIFT features for scene A.

In this experiment, the scene was captured at 100 cm distance from the camera. Objects were captured separately from different viewpoints under varying illumination and geometrical conditions. This procedure, being essential for the system's training was done while the system was offline. Objects are stored as images into the database for further exploitation. According to the first stage of our algorithm the SIFT features of all images in the database should be extracted. In Table 1(a), the results of this phase for the scene A are presented. Apparently, more features are extracted from the most textured objects. The ones containing high amount of local extend favor during the SIFT feature extraction procedure.

In Figure 1(b) the extracted SIFT features from the scene A and the contained objects are depicted. The second stage of our algorithm involves the matching sub-procedure of SIFT, where the features of the scene under investigation and

Table 1. a) SIFT features for scene A and its contents. b) The results of the position estimation procedure for scene A.

(a)

Image	SIFT features
Scene A	208
Book	150
Box	349
Modem	108

(b)

Object	Matches	Z	d_0	d_S	Z^*
Book	18	100	70	1.4992	104.93
Modem	20	120	70	1.6098	112.68
Box	44	140	70	2.005	140.65

of each separate object are compared to obtain the N matches. The scene is collated with all possible objects' views. The total amount of these comparisons depends on the quantity of different views of the same object. During the database construction we altered objects' viewpoint and distance from the camera. It is apparent that, proposed method's efficiency is directly related to the size of the constructed database. A large database maximizes the possibility to perform adequately recognition and more accurate position estimation tasks. The results of these comparisons are shown in the second column of Table 1(b). As it was expected, the object "box" provided more matches. This is due to the fact that it outperformed during the feature extraction process. The most essential issue that one should keep in mind is that, there is a direct relationship between the amount of SIFT features extracted and the final amount of matches. The more the SIFT features are the more the object's matches with the scene and as a result, the higher the possibility the object to be found in the scene.

The next step includes the spatial data estimation. In the third column of Table 1(b) distances in cm for every object in scene A are shown. By applying the last stages of our method we are able to estimate d_S. In particular, we use the pre-computed depth d_0 which is object's distance from the camera when the later is captured alone during the training session. Pre-computed depth d_0 is definitely different from object's distance from the camera (Z) as it is shown in Table 1(b). In order to confirm proposed method's accuracy we needed a ground truth. For this purpose when the scene is captured, we measure and store objects' distances from the camera (Z) with a laser distance measuring device. The goal of the proposed method is to approximate the later distances with as high accuracy as possible. In particular, by executing the final stages of the algorithm we are able to estimate objects' distance (Z^*) from the camera. In the last column of Table 1(b) the results of the position estimation process are illustrated. Figure 2(a) depicts the results of the proposed method. More specifically, in white boxes the number of N matches between the scene and the object is shown. Moreover, in the same figure, the mean Euclidean distances E_S and E_0 extracted in Stage IV, are illustrated. In addition, ratio d_S for every object in a scene is extracted. The above are illustrated as a modified MATLAB figure title that is constructed through the proposed algorithm.

In order to assess proposed method's robustness under 3D rotation and illumination changes we took measurements on the scene shown in Figure 2(b). The

(a) (b)

Fig. 2. (a) Proposed method's results for every object in scene A. (b) Scene B containing rotated objects.

new scene contains the same objects as the previous one but in different alignment. Especially, objects were rotated whilst keeping the same distance (Z) from the camera and altering the illumination conditions. In Table 2(a), the results of the first stage of the proposed method for scene B and its contents are presented. In addition, in Table 2(b) the results of the proposed position estimation process are illustrated. Although scene B differs significantly from A, the efficiency of the proposed object recognition and position estimation method, remains high. In addition to the above results, we introduce an efficiency ratio or an accuracy percentage with a view to further analysis. Thus, we estimate:

$$\alpha = (1 - \frac{\| Z - Z^* \|}{Z}) \times 100.$$

Table 2. a) SIFT features for scene B and its contents. b) The results of the position estimation procedure for scene B.

(a)

Image	SIFT features
Scene B	285
Book	150
Box	349
Modem	108

(b)

Object	Matches	Z	d_0	d_S	Z^*
Book	26	100	70	1.410	98.756
Modem	22	120	70	1.698	118.86
Box	42	140	70	1.957	137.00

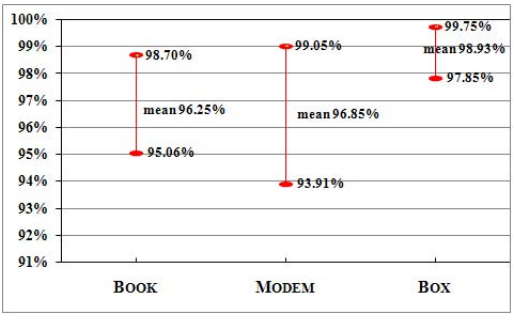

Fig. 3. Accuracy of the proposed algorithm

The proposed algorithm was evaluated through 30 tests and the results acquired are illustrated in Figure 3. In these experiments 3D rotation (up to 30 degrees) and illumination changes were applied. In Figure 3 minimum, maximum and mean accuracy values are respectively shown. Moreover, the most interesting issue that derives is the fact that the extracted accuracy never drops below 93%. Furthermore, as expected larger objects are found with a higher accuracy.

5 Conclusions

A technique for object recognition based on SIFT has been presented in this paper. The novelty here is that we have enriched it to perform depth estimation and, consequently, object localization in an arbitrary scene. The key idea is that the tracked SIFT features are located on given geometric positions, thus they can be considered as the corners of a polyhedron, the center of gravity of which is computed and it is associated to the actual center of mass of the sought object. Once the features' center of mass is known and the object is recognized, the distance of the object from the camera is trivial, given at least one recorded position of the object. The application of such a method in robotic manipulation tasks is apparent, as it permits the identification of the object to be handled, as well as its distance from the robot's eye. Experimental results, where different illumination conditions and objects' viewpoints were sampled, prove that larger objects are easier to be found, whilst, their position is estimated with higher accuracy. To sum up, with an outlook to future work, we will enrich and modify the algorithm in order to perform challenging pose estimation tasks. The overall purpose of the new framework will be the extraction of objects' roll, pitch yaw angles and T_X, T_Y and T_Z translation matrices. Thus, several tasks such as automatic manipulation of objects or assisting autonomous vehicles avoiding obstacles would be completed at high execution times with significant efficiency.

Acknowledgements. This work is supported by the E.C. under the FP6 research project for Autonomous Collaborative Robots to Swing and Work in Everyday EnviRonment ACROBOTER, FP6-IST-2006-045530.
http://www.acroboter-project.org

References

1. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR 2006: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp. 2161–2168. IEEE Computer Society, Los Alamitos (2006)
2. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV 2003: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA. IEEE Computer Society, Los Alamitos (2003)
3. Schleicher, D., Bergasa, L., Barea, R., Lopez, E., Ocana, M., Nuevo, J.: Real-time wide-angle stereo visual slam on large environments using sift features correction, October 29 - November 2, pp. 3878–3883 (2007)
4. Davison, A.J., Molton, N.D.: Monoslam: Real-time single camera slam. IEEE Trans. Pattern Anal. Mach. Intell. 29(6), 1052–1067 (2007); Member-Reid, I.D., Member-Stasse, O.
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
6. Hoshi, Y., Kobayashi, Y., Kasuya, T., Fueki, M., Kuno, Y.: Interactively instructing a guide robot through a network. In: International Conference on Control, Automation and Systems, 2008. ICCAS 2008, pp. 1841–1845 (2008)
7. Zhao, L., Li, R., Zang, T., Sun, L., Fan, X.: A Method of Landmark Visual Tracking for Mobile Robot. In: Xiong, C.-H., Liu, H., Huang, Y., Xiong, Y.L. (eds.) ICIRA 2008. LNCS (LNAI), vol. 5314, pp. 901–910. Springer, Heidelberg (2008)
8. Liao, M., Wei, L., Chen, W.: A novel affine invariant feature extraction for optical recognition, vol. 3, pp. 1769–1773 (August 2007)
9. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 27(10), 1615–1630 (2005)
10. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22(10), 761–767 (2004)
11. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. 110(3), 346–359 (2008)
12. Lindeberg, T.: Scale-space theory: A basic tool for analysing structures at different scales. Journal of applied statistics 21(2), 414–431 (1994)
13. Meger, D., Forssen, P., Lai, K., Helmer, S., McCann, S., Southey, T., Baumann, M., Little, J., Lowe, D., Dow, B.: Curious george: An attentive semantic robot. Robotics and Autonomous Systems 56(6), 503–511 (2008)
14. Forssen, P.E., Meger, D., Lai, K., Helmer, S., Little, J.J., Lowe, D.G.: Informed visual search: Combining attention and object recognition, pp. 935–942 (2008)