Monocular Camera based Object Recognition and 3D-Localization for Robotic Grasping

Aashna Sharma, Ishant Wadhwa, Rahul Kala
Robotics and Artificial Intelligence Laboratory
Indian Institute of Information Technology Allahabad
Allahabad, 211012, India
{aashnashrm.19,ishant.wadhwa5, rkala001}@gmail.com

Citation: A. Sharma, I. Wadhwa and R. Kala (2015) Monocular camera based object recognition and 3D-localization for robotic grasping. In *Proceedings of the 2015 International Conference on Signal Processing, Computing and Control*, Waknaghat, pp. 225-229.
Final Version Available at: http://ieeexplore.ieee.org/document/7375030/

***Abstract*—Visual Servoing for a robotic hand is still a difficult problem to solve and is the topic of current research. To recognize an object of interest and calculating its orientation and location without extensive training is another far-fetched problem faced by researchers in this field. In this paper, object recognition based on scale-invariant feature based transform (SIFT) is done, which is used to calculate the location in image plane, scaling factor and orientation of the object in the testing image with respect to the training image of the object. Here the system is trained using single image instead of many images at different orientations. Further, the object location from 2D image coordinated is mapped to real world 3D coordinates where the third dimension is calculated using relative scaling obtained from SIFT. The images are obtained using a stationary monocular camera system. For this, we have used *NAO* which is autonomous, programmable humanoid robot developed by Aldebaran Robotics, France.**

Keywords-Visual Servoing; SIFT ; Monocular camera system; humanoid robot

I. INTRODUCTION

Many robots have been developed by far now which offer assistance to people in completing various day to day tasks like taking care of our families or interacting with children and helping them in learning as well as recreational activities. For the robots that are used in industries and at homes, grasping or manipulating an object is a task which requires us to calibrate the relationship between the object's position and robot's arm's position in order to manipulate the object accurately (i.e. for structured environment).

For a robot to grasp an object, it needs to know the orientation, pose and location of the object apropos to its base frame with respect to which its end effector's position is defined. Once we get both with respect to the same frame of reference then we can easily manipulate the object with robot's hand. For an object to be manipulated in an unstructured environment visual servoing can be used to direct the robot to the object of interest [1].

A lot of comprehensive study has been done in the past for feature detection among which Harris corner detector [2], ShiTomasi features [3], SIFT features [4], and Maximally Stable Extremal Regions [5] are most popular methods. The robustness of an object recognition and localization system depends on the successful extraction of ample features for each object. It has been shown that by using a calibrated stereo system, we can achieve higher accuracy [6]. However, for the ease of computation and simplicity, a monocular camera system and feature correspondences based localization is performed [7]. Therefore, on the basis of scaling, depth information is determined. The SIFT method provides the most precise results of feature detection and feature description. SIFT [4] can robustly identify objects even among clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform scaling, orientation and partially invariant to affine distortion and illumination changes [8]. SIFT is robust in extracting features from typical images and can efficiently locate small objects in cluttered environment. in this paper, we have used SIFT features to find correspondence between objects and to locate them in 2D image plane.

In this work, we suggest a framework to recognize and localize the object in 3D using monocular camera system of NAO robot Fig. 1. We create a database of objects with known

Fig. 1 : NAO Robot.

geometry, train the system with single image of each object and then, test the system with images of any of the object in an unknown environment. The location and orientation of the object with respect to its training image is obtained by comparing the SIFT features of test image to those of training images, thereby obtaining the perfect match. In section II, we discuss the methodology, the process where SIFT is used for object recognition and 2D localization of the object in image plane and localization of the object in 3D real world coordinates.Section III gives the experimental results and conclusion remarks are given in Section IV.

II. METHODOLOGY

The flowchart of the methodology used is shown in Fig.2

A. Object recognition based on SIFT

1) Extracting SIFT keypoints

In this section, we introduce a brief method for extracting keypoints of an image using SIFT.

First, we create a scale space of the images. For this, we magnify the original image 2 times and then downsample the image with scale 2. This process is continued till an image of size 64 × 64 is obtained. Construct a Gaussian pyramid and search for local extrema (peak points) in difference of Gaussian (Do G) at each octave of the pyramid.

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

$$L(x,y,\sigma) = G(x,y,\sigma) \times I(x,y) \quad (2)$$

$$D(x,y,z) = L(x,y,K\sigma) - L(x,y,\sigma) \quad (3)$$

$L(x,y,\sigma)$, denotes the scale space of an image which is obtained by convolving the image, $I(x,y)$, with the variable scale Gaussian kernel, $G(x,y,\sigma)$. SIFT keypoint locations are computed from, $D(x,y,\sigma)$, the difference-of-Gaussians with a multiplicative constant, K.

Characterization of the image at all the key locations is done by extracting the gradients and orientations of image at each level of pyramid of the smoothed image. At each point, $L(x, y)$, the magnitude of image gradient, $M(x,y)$ and orientation, $\theta(x,y)$ is found out by the equations (4) and (5):

$$M(x,y) = \sqrt{\left(L(x,y+1) - L(x,y-1)\right)^2 + \left(L(x+1,y) - L(x-1,y)\right)^2} \quad (4)$$

$$\theta(x,y) = arc\ tan \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)} \quad (5)$$

Since we are using *arctan* function, there may be multiple values of orientation, out of which one will be the actual orientation.

Now, for keypoint description, consider a 16 × 16 window around each keypoint which is further divided into 16 segments of 4 × 4 window. For each sub-block, calculate gradient magnitude and orientation of each keypoint. By this, we obtain 16 directions and out of these 16 directions calculate orientation histogram for 8 directions starting from 0° to 360° with an increment of 45° at each level i.e. 0°, 45°, 90°, so on. Thus, each keypoint has 4 × 4 × 8=128 descriptors.

2) Matching algorithm and Object localization

To find the closest match for the testing image find the SIFT keypoints as discussed in step II.A.1. We find the closest match only if the Euclidean distance $d(D_1, D_2)$ between the descriptor, D_1, and the descriptor, D_2, multiplied with the threshold is less than the distance of descriptor, D_1, to all the other descriptors. In this process, many inconsistent matches also get selected and to remove these inconsistent matches, we use Hough transform [9].

Considering a space of all possible poses, Hough transform is used for pose clustering in which the potential matches cast vote for poses. Even a single matching feature can cast its vote and the pose with the maximum number of votes has the highest probability of being correct. We use these matches with highest number of votes to get a more accurate transformation. Here, Hough transform is used as it is cheap to compute and provides accurate results for best matches by reducing the number of futile votes and removing the invalid matches.

To find scale and rotation, the consistent matches are used as follows:

For the image $I_i(x_i, y_i, a_i, s_i)$ *where* $i=1,2$

Let, $v_i = (-x_i, -y_i)^T$

$$sr = s_i / s_{i+1} \quad (6)$$

$$da = a_i - a_{i+1} \quad (7)$$

Then, $v_{i+1} = R \times (v_i / sr)$ (8)

Where,

$$R = \begin{bmatrix} \cos(da) & \sin(da) \\ -\sin(da) & \cos(da) \end{bmatrix} \quad (9)$$

where, v_i is the vector of centre offset of the image, sr is scale ratio and da is the difference in angles of the two images and R is rotation matrix.

B. Monocular camera based 3D object localization

After detection of the object in 2D image plane, the location of object in 3D can be estimated.

1) Camera Calibration

Camera calibration is the process of calculating the intrinsic and extrinsic parameters of a camera. The intrinsic parameters of camera include focal length, principle point and skew coefficient which denote the transformation from 3D camera coordinates to the 2D image coordinates. The extrinsic parameters are rotation and translation which denotes the transformations from 3D world coordinates to 3D camera coordinates. In the present study, chess board camera calibration has been used [10].

To accomplish this task a chess board is printed with fixed box size, here we have taken 5 × 7 chess board with boxes of size 30 mm on an A4 size paper. It should be mounted on a smooth surface like a table top or marble tile. Before clicking images from the camera it should be made sure that the focal length of the camera is constant. Click at least 8-10 pictures from the camera and consider at least 6-8 which are of good quality for calibration process. With these images we calculate intrinsic parameters of the camera and hence derive intrinsic matrix M_{int} i.e.

$$M_{int} = \begin{bmatrix} f_1 & a \times f_1 & c_1 \\ 0 & f_2 & c_2 \\ 0 & 0 & 1 \end{bmatrix} \quad (10)$$

Where f_1 and f_2 are focal length of the camera, α is the skew and c_1 and c_2 are principle points of the camera.

2) 2D to 3D mapping using monocular camera system

Step 1: Using the position of object recognized in image plane as,(u_i, v_i), we know that,

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = M_{int} \times \vec{x} \quad (11)$$

$$\therefore \vec{x} = [M_{int}]^{-1} \times \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} \quad (12)$$

Here $\vec{x}$ is the 3D camera coordinate derived from 2D image plane with its third coordinate as 1.

Step 2: To compute the third coordinate in the camera frame, focal length of the camera f is multiplied with the scaling factor s found out from section *A.2*,

$$X_{c,3} = f \times s \quad (13)$$

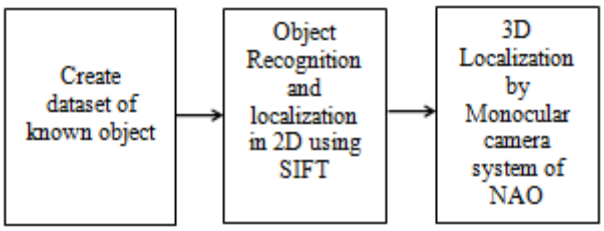


Fig.2 :Flowchart of method used.

Hence we compute the other two coordinates in the camera coordinate frame as,

$$\vec{X}_c = \frac{\vec{x} \times \vec{X}_{c,3}}{f} \quad (14)$$

Step 3: Now taking in consideration the extrinsic parameters (rotation R and translation t) of the camera, compute the real world coordinates $\vec{X}_w$ of the object as,

$$\vec{X}_w = [\vec{X}_c - t] \times R^T \quad (15)$$

This process dependent on relative scaling and will provide us with the approximate position of the object in real world.

III. EXPERIMENTAL RESULTS

Here we realized a realistic scenario where a known object in Fig. 3 (a) was recognized in an unknown environment i.e. in Fig. 3 (b) and located in 2D image plane, Fig. 7. In Fig. 4 SIFT features of the image were found out, described by yellow circles. The total number of features detected in testing image were 900 which is an ample amount required to describe an image. We recognized the object from the database by matching these features on the basis of total number of features matching and the distance between the features matched. Fig. 5 shows the object recognized and all the matching features between the two images. Out of these matches some were inconsistent which are not required and we need to remove them. According to Hough transform, we plot between center offset coordinates x versus y of the image as shown in Fig. 7, and scale versus angle of the image as shown in Fig. 6, to find out the actual orientation of the object recognized on the basis of maximum clustering in the plot. This removed all the inconsistent matches as well as found out the actual orientation and scale of the object with respect to its training image. Fig. 8 shows all the consistent matches. Fig.9 (a) shows the object of interest segmented out from the unknown environment. In Fig. 9 (b), the object is localized in 2D image plane with its orientation defined. Table I shows the residual error in pixel values which depicts the correctness of matches found out.

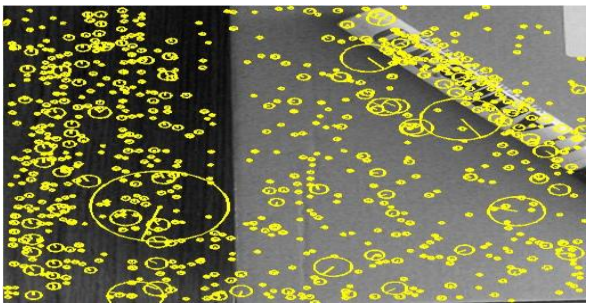


Fig. 4 : SIFT features detected in testing image.

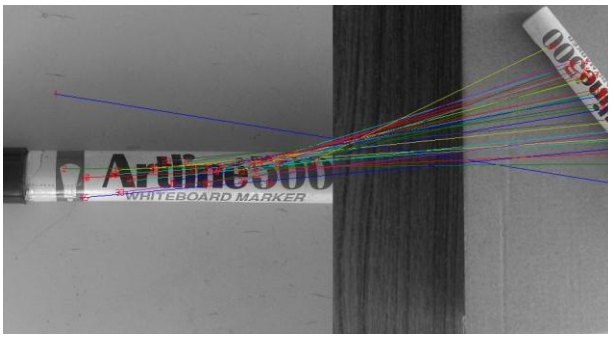


Fig 5 : Matching features of testing image with training image.

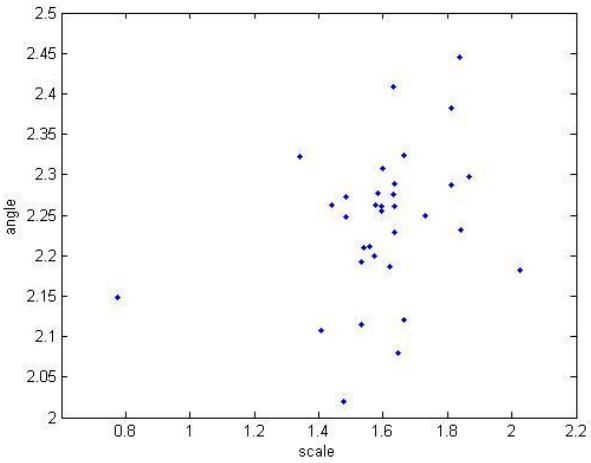


Fig. 6 : Plot between scale and orientation of the image pixels

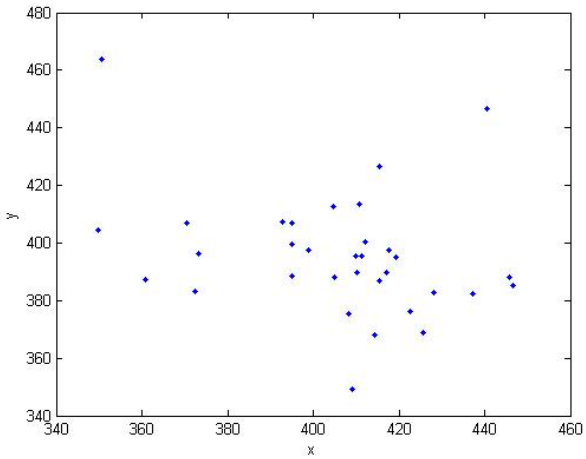


Fig. 7 : Plot between x and y coordinates of the center offset of image pixels.

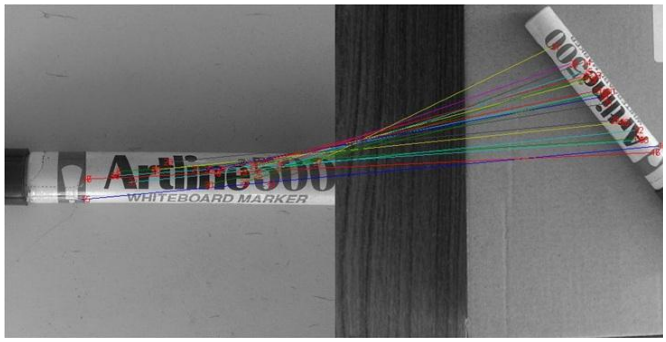


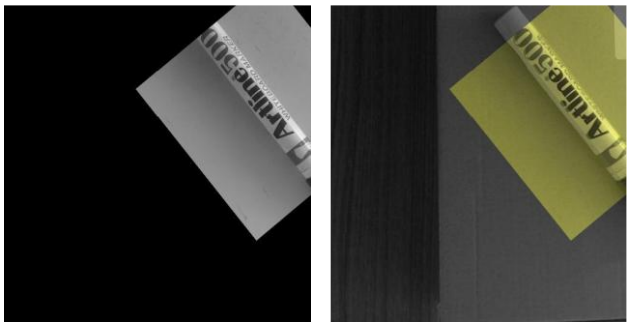
Fig.8:Relevant Matching Features of training and testing images.

Then we calibrated NAO's camera with the help of chess board of 5 × 7 boxes. Fig. 10 shows the corners detected of the chess board with green colored lines represent X and Y axis in the image plane of chess board and red crosses show the corners detected in the image. Hence we computed the intrinsic parameters of the camera f NAO, given in Table II. Fig. 11 shows the plot of the image plane with respect to the camera

(a) (b)

Fig. 3 : (a) Training image.(b)Testing image.

Skew	alpha_c = [0.00000] ± [0.00000] => angle of pixel axes = 90.00000 ± 0.00000 degrees	
		Points
1	7.666000	115



(a) (b)

Fig. 9: (a) Segmented out object of interest.
(b) Object of interest overlapped on testing image.

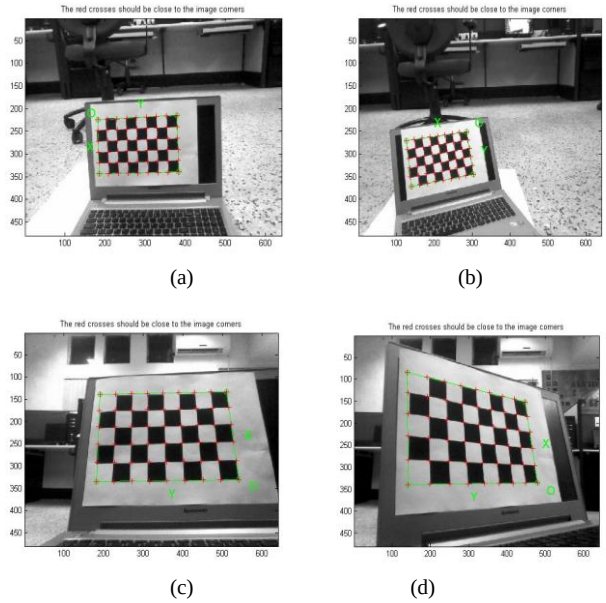


Fig.10 : Corner detection in camera calibration.

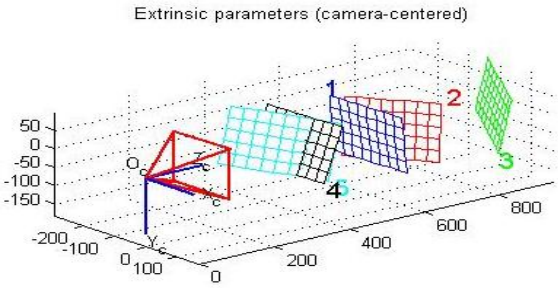


Fig. 11 : Extrinsic parameters.

coordinate plane i.e. extrinsic parameters of the images used for calibration process. The error between the image plane coordinates and real world coordinates according to the intrinsic and extrinsic parameters of the camera was calculated to be 0.0667.

IV. CONCLUSION

In this demonstration, we realized two main modules of Monocular camera based Object Recognition and 3D-Localization for Robotic Grasping system. First, we recognized a known object from our object database, located it in 2D image frame and then localized it in 3D world coordinate frame. By the use of monocular camera system, computational expensiveness of the system is reduced and hence, can be realized in near real time. Along with that, we computed 3D positions of the object in real world with high accuracy. Using SIFT algorithm for object recognition and localization in 2D, even textured objects were easily recognized and localized in complex unknown environments. In addition, we also calculated the change in orientation of the object recognized with respect to its training image. In near future, we will compute the coordinates of the object with respect to base frame of NAO, solve inverse kinematics for end effector of NAO and hence realize grasping of the recognized object.

ACKNOWLEDGMENT

We would like to thank Mukul Anand Bisherwal, a fellow student at IIIT Allahabad for helping us in editing the paper. This work is supported by the Indian Institute of Information Technology Allahabad.

REFERENCES

[1] A. Nakhaei and F. Lamiraux, “Motion Planning for Humanoid Robots in Environments modeled by Vision,” IEEE Int. Conf. on Humanoid Robots, pp. 197–204, 2008.

[2] C. Harris and M. Stephens, “A Combined Corner and Edge Detector,” Proc. of the 4th Alvey Vision Conf., pp. 147–151, 1988.

[3] J. Shi and C. Tomasi, “Good Features to Track,” IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 593–600, 1994.

[4] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” Int. Journal of Computer Vision, Vol. 60, No. 2, pp. 91–110, 2004.

[5] S. Obdrzalek and J. Matas, “Object Recognition using Local Affine Frames on Distinguished Regions,” in Proc. British Machine Vision Conf., Vol. 1, pp. 113–122, 2002.

[6] Chia-Hung Chen; Han-Pang Huang; Sheng-Yen Lo, "Stereo-based 3D localization for grasping known objects with a robotic arm system," Intelligent Control and Automation (WCICA), 2011 9th World Congress on , vol., no., pp.309,314, 21-25 June 2011.

[7] Bodo Rosenhahn. "Foundations about 2D-3D Pose Estimation". CV Online. Retrieved 2008-06-09.

[8] Lowe, David G. (1999). "Object recognition from local scale-invariant features". Proc. of the International Conf., on Computer Vision 2. pp. 1150–1157.

[9] D.H. Ballard, "Generalizing the Hough Transform to Detect Arbitrary Shapes", Pattern Recognition, Vol.13, No.2, p.111-122, 1981.

[10] Heikkila, J.; Silven, O., "A four-step camera calibration procedure with implicit image correction," Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on , vol., no., pp.1106,1112, 17-19 Jun 1997.