A Review of Object Recognition and Localization Methods with SIFT based Approach

Vaibhav babrekar
Dept. of EXTC,
PRMIT&R, Badnera
Amravati, India
Vaibhav.babrekar@gmail.com

Girish Patil
Dept. of EXTC,
PRMIT&R, Badnera
Amravati, India
patilgirish213@gmail.com

Ashay Rokade
Dept. of EXTC,
PRMIT&R, Badnera
Amravati, India
rokadeashay@gmail.com

***Abstract*— In recent years, both academia and industry is focusing significantly on Vision based pick and place robotic systems. The system typically employs machine vision to analyse the scene, identify and locate the specified object and provide feedback to the robot arm for subsequent operations. For efficient and successful picking, the vision system needs to identify the exact position and the orientation of the objects, the Scale Invariant Feature Transform (SIFT) is used for this purpose. The concept of the proposed work is developed around two major areas; object recognition for developing artificial vision system and the robotics for carrying out the specified task with the specified object. In this paper, such object recognition techniques are reviewed.**

Keywords*— Scale Invariant Feature Transform, Image Processing, Object Recognition, Robotics, Feature Extraction

I. INTRODUCTION

The pick-and-place processes are the primary requisite for many of the industrial and household application. For such applications, there is a need to automate the pick-and-place process basically comprising of picking the intended objects, possibly performing certain tasks and placing them to desired location. The automated pick-and-place systems mainly consist of robotic arms and sensors. The machine vision is used as sensor and the primary function of them is to drive the robotic arms to the right location of desired object for picking and placing according to the robot's degrees of freedom. The placing location is prefixed in most applications hence the sensors are rarely used here and the placing phase found to be comparatively easier. In contrast, the picking phase becomes very complex process in the applications where the scene is occluded and constrained. In this phase the sensors plays most important role as it is responsible for correct movements of the robotic system. [1]

Most of the picking system assumes the situations where the objects are well structured, ordered, aligned and synchronized grasping of the objects. For such cases, the use of simple photocells will be sufficient to accomplish the picking phase. However, this approach will not be adequate for several applications as the arrangement to keep the objects well-structured and well aligned results in wastage of time and space of the process. In addition to this, there are some applications where the objects need to be kept in bins for saving time and/or for hygienic and safety reasons as shown in Fig.1. In this case, cameras used must be high resolution along with appropriate machine vision algorithms. In the scene, the objects are positioned at random inside a bin, container or even at random on a belt/shelf; this problem is addressed as bin picking. [2]

The pick-and-place systems with robotic vision present several challenges, like the system should be capable of overcoming the difficulties in the disposal of the objects such as order, structure and placing of grasping points, working with every type of object of different dimension and complexity, with reflective surfaces or semi-transparent parts, such as in the case of pharmaceutical and cosmetic objects, often reflective or included in transparent flow packs, tackling the conditions of occlusion and clutter which make the object only partially visible, not only counting but also classifying the first instance of the object and also to identify all the duplicate objects with their orientation and dimensions, meeting the required working speed though having fast detection technique so that it works with several objects per minute.

The proposed approach can meet all these challenges by proposing a feature-based segmentation technique i.e. SIFT able to segment multiple occluded objects. Moreover, this has been a very active area of research in the last decades and as indicated by the tremendous amount of work and documentation published around this. As needs change and become more demanding, researches have been encouraged to develop new technologies in order to fulfill these needs. In this tenor, it is worth mentioning that many methods published satisfy the everyday needs of pick and place system including feature detection, matching and 3D modeling. Recognizing the correct location and pose of the given object is the ultimate purpose of robotic vision system. More than a decade ago, the applications associated with 3D models and object reconstruction were mainly for the purpose of visual inspection and robotics.

In this paper, several object recognition and localization techniques used for implementing automated pick and place systems with main focus on SIFT based technique are reviewed.

The organization of this paper is as follows. The next section presents the literature review of various object recognition methods from both technological and application perspective. Section 3 will briefly describe the overall system approach for object detection with SIFT algorithm. At last, Section 4 summarizes the contributions of the paper and draws the conclusion.

II. Literature Review

A very luxuriant literature related to robotic vision for pick-and-place systems is available. Object recognition and localization are the basic processes of the robotic vision. A wide range of information composed of various techniques applied for the object recognition and localization is known. According to application perspective, the initiating work utilized the basic image processing techniques such as thresholding, segmentation to identify and to locate the grasping point which was the center of gravity for that object.

Rahardja and Kasaka developed a vision algorithm that provides adequate information for a bin picking robot to handle complex industrial objects such as alternator covers as the target objects. This algorithm is capable of detecting such complex objects as well as to estimate their 3D pose by stereo vision technique. In this work, the landmark features are defined as seed features i.e. unique and easily detectable and supporting features that assist both identification and pose estimation purpose. In short, vision algorithm works as follows:

Initially, the left and right stereo images are provided to system and then regions of interests i.e. areas which might contain the landmark features, are extracted to give initial estimate of visual cues by feature detector/predictor (FDP) module. After this stage, 2D appearance is evaluated from which feature verification (FV) module examine and decide whether to accept or reject each given estimate. As the verification is over, the stereo correspondence and pose estimation (SCPE) module checks for the matches of estimates from both views and determine the object pose. Finally, The manipulator interface(MI) module provide the appropriate commands for relevant operation of manipulator, for each identification and pose estimation pair result. This approach succeeds in minimizing the computational complexity by using simple features of industrial objects. But, a great deal of human assistance is required to construct the valid choice of object model, in order to make this system feasible solution to given task and very simple image processing techniques are followed here which cannot be applied in intricate scenes consisting of multiple objects. [2]

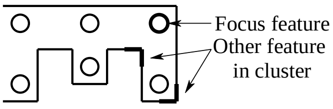

Fig.1: An example object with feature cluster

An alternative method is the detection-by-feature approach which looks for discriminative features which are unique for that object i.e. local feature. This proposal is discussed in [3]; it is called local-feature-focus. This algorithm can recognize and locate partially visible 2D objects, by performing the segmentation on higher-level features, such as round holes and convex or concave 90° corners; instead at pixel level.

The position, size and orientation of each such feature are registered. The algorithm looks for a cluster of local features in a relative configuration which defines the characteristic specific to that object. One feature in the cluster is chosen as the “focus” feature, i.e. the one with respect to which the other features are located. This focus feature is searched first, while searching for the object. This approach also explicates complex structure of features, by means of feature indexed hypotheses and binary decision trees. These methods exploited very specific features that are limited for the particular object only, hence cannot be extended for whichever type of object.[4]

Color provides powerful information for object recognition. Numerous methods exists that uses color as basis for object recognition. But these methods not showed effective robustness against the changing illumination across the scene, changing geometry of object, object occlusion and cluttering. One of the researcher proposed the color based segmentation. The purpose of the research by Gevers. And Smeulders is to achieve the recognition of multi-colored objects invariant to most of the above stated constraints. Assuming dichromatic reflectance and white illumination, some new color models are presented alongwith the existing one i.e. normalized color rgb, saturation S and hue H, and the newly proposed color models c1 c2 c3 and l1 l2 l3. Extensive experiments are carried out for testing of these color models against the various imaging conditions and it is shown that all are invariant to a change in viewing direction, object geometry and illumination. Unfortunately, the clutter of object appearance due to occlusions and distractors results in making these approaches unreliable in real scenarios.[5]

An alternative approach is the use of edges to compute the object boundaries and shape. Often the edge information is fused with region-based information which provides the strong basis for object recognition. Mueller et al. presents technique where the edge-based initial segmentation provides the initialization for a region growing algorithm. Edge-based methods are able to detect long, straight edges while region-based techniques used to close gaps within these edges. Region-based approaches exploit the homogeneity of objects while uncertainties in detecting the exact boundary positions can be reduced by previously extracted edges. This approach is a combination of region and edge based segmentation techniques that incorporate new methods for the evaluation of the straight edges and edge guided region growing.[6]

The previous method is very impressive and proved excellent accuracy in several contexts. But it is sensitive to occlusions and reflexes. A challenging job is to develop recognition scheme for free-form objects independent of viewpoint, clutter and occlusions. In [6], a novel 3D model-based algorithm is presented as a perfect solution to obtain the desired scheme which works automatically and efficiently. This algorithm carry out two phases of operation i. e. offline and online. In offline phase, multiple unordered range images (views) are captured and 3 D model of an object is automatically constructed. These views are transformed into multidimensional table representations referred as tensors. The tensors of a current view and those of the remaining views are by simultaneously matched using a hash table-based voting scheme to establish the correspondences between various views. As a result of this, a graph of relative transformations is obtained which is used to register the views prior to integration into a coherent 3D model. The model library constitutes these models and their tensor representations. In online phase, the tensors of scene are computed and simultaneously matched with those in the library by casting votes. The model tensor receiving most votes said to have highest similarity is then transformed to the scene. The alignment measure is performed, if the model perfectly aligns with an object in the scene, then it can be adjudged that the object is recognized. The same process is repeated until the complete segmentation of scene. The exclusive experiments have shown that, this scheme is very efficient for automatic 3D object recognition and segmentation in the presence of clutter and occlusions. But the average processing time per image is in the order of tens of seconds, hence it is less applicable to real time object recognition.[6]

Agrawal et al. presented a system based on using depth edges (silhouettes) of objects. This system employs a novel vision sensor comprising of a camera surrounded by eight flash lights. The images under different flashes are captured and their shadows are observed to obtain the depth edges or silhouettes in the scene. The silhouettes of different objects are segmented and each silhouette is then matched with object silhouettes in different poses stored in database to estimate the coarse 3D pose. A Computer Aided Design (CAD) model of the object is used to compute the database. A fully projective formulation of Lowe's model based pose estimation algorithm has been used to refine the pose. The estimated pose is assigned to coordinate system of robot utilizing the hand-eye and camera calibration parameters, which permits the robot to pick the object.

This system can handle complex ambient illumination conditions, challenging specular backgrounds, diffuse, reflective and texture-less objects with high working speed though simple and fast for practical implementation. This system has some limitation as it cannot handle stacked specular object thin objects. The dark background produces difficulties as it reduces the contrast. Moreover, transparent and translucent objects cannot be handled by this technique.[6]

One of the author fused the vision-based techniques applied to the 3-D data obtained using range sensors and Radio Frequency Identification (RIFD) technology. As the 3D vision-based algorithms having some limitation in practical applications such as uncertainty, incapability for real time operation but proved acceptable for pose estimation. Whereas RFID technology has potential to detect the presence of the object in given occluded scene but appeared ineffectual in proper localization of it. This paper presents a powerful and an effective machine vision technology that coalesces both i.e. 3D vision-based algorithms and RIFD technology to detect the ubiquitous objects which are solid in nature and have well-defined geometrical models.[7]

For the implementation of this technique, an RIFD tag is needed to attach on every object in scene. Once the attached tag is detected, the geometric model for that particular object is extracted from the local database and its pose is estimated. In general object recognition methods employing a 3D vision-based algorithm, the 3D models are extracted from the current image objects and then it is compared with the models of the objects from the entire database. As a result of such matching, the reduced set of objects is obtained having higher probability to be present in the current scene. The objects with higher matching score possess higher probability to be present in the scene. Normally, the iterative closest point (ICP) algorithm is exercised to work out the matching process. The complexity of computation is more in this process. With the additional information from RIFD tags, the searching tree becomes smaller and the false matches with similar objects also mitigates. This approach is suitable for medium and large databases. Moreover, the extra labour is needed to attach the tag to each object, thus this approach is somewhat tedious and time consuming at initial development.[8]

Another author has used the similar approach, in place of RIFD tags; the Quick Response (QR) codes are utilized. This QR codes are pasted on each object and the object is recognized by recognizing the QR code. But this method also involves lot of human labour in pasting the QR codes.[9]

As the approaches stated above has the limitations and cannot tackle the required challenges satisfactorily, there is a need of a suitable approach which uses a point-based or feature-based solution. In this category, single features are matched between a model of the target object and the current image and these matches are used to identify and locate the objects. Based on features, the Dominant Orientation Templates (DOT) and HoG (Histograms of Gradients) are some interesting approaches.DOT calculates local gradients, it depends on locally dominant orientations and this approach is explicitly invariant to small deformations and translations. Though these approaches give the impressive results, they result in performance degradation in case of occlusions. [1]

Mikolajczyk and Schmid presented a comparison of several feature descriptors practiced for object detection such as shape context, steerable filters, PCA-SIFT, differential invariants, spin images, SIFT, complex filters, moment invariants and cross-correlation. These descriptors are computed for local interest region extracted by five different detectors. To accomplish this result, the descriptors are evaluated on real images with different geometric and photometric transformations such as rotation, scale change, viewpoint change, image blur, JPEG compression, and illumination. They concluded that SIFT descriptor outperforms the others independently and emerges as the best method for object recognition and localization suitable for pick and place robotic systems.[10]

The SURF based algorithm used in [11] recognizes objects with invariant features, and reduces dimensions of the feature descriptor to reduce computational time. The experimental result shows that their work was fast and robust than the traditional methods and can track objects accurately in various environments.

III. Proposed System

The proposed approach consists of two steps:

1. Extraction and matching of features: Number of significant features is extracted from the object model as well as from the current image; then by applying a proper similarity measure, these features are matched and the best correspondences are kept for next phase;

2. Localization of object: From the obtained set of correspondences, registration transform between the model is computed and the best location of the detected object in the current image.

The first step can be further divided in three sub-steps.

- The off-line extraction of features on the model (Fig.3)
- The on-line feature extraction on the current image (Fig.4)
- The matching between the model and the image

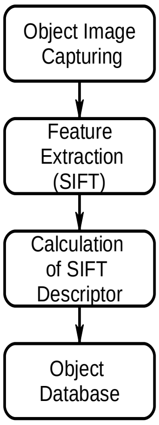

Fig.3: The off-line extraction of features on the objects

Among many methods for local feature extraction and model matching, here the SIFT and Two Nearest Neighbour (2NN) is selected as SIFT has proved to be very robust to noise and invariant to scaling, rotation, translation and (at some extent) illumination changes as well as compatible for real time applications.

The SIFT can be used to obtain the set of keypoints KM for model M,

$$K_M = \{k_i \triangleq [x_i^M, y_i^M, \theta_i^M D_i^M], \quad i = 1,2,\dots p\}\text{.......(1)}$$

Where, where x and y are the 2D image coordinates, D the 128-value SIFT descriptor and θ the main orientation computed by SIFT. After this the system is applied on-line to current image I, resulting in set of keypoints KI ,

$$K_I = \{k_j \triangleq [x_j^I, y_j^I, \theta_j^I, D_j^I], \quad j = 1,2,\dots q\}\text{...........(2)}$$

From the two sets KM and KI, the standard 2 Nearest Nieghbour algorithm computes the Euclidean distance between $D_i^M$and D_j^I to determine the corresponding model to image matches $M = (m_1, m_2, \dots, m_N)$ where each match mq contains the (x,y) coordinates on the two reference systems and the main orientation on the current image $m_q = \{(x_j^I, y_j^I, \theta_j^I), (x_i^M, y_i^M)\}$

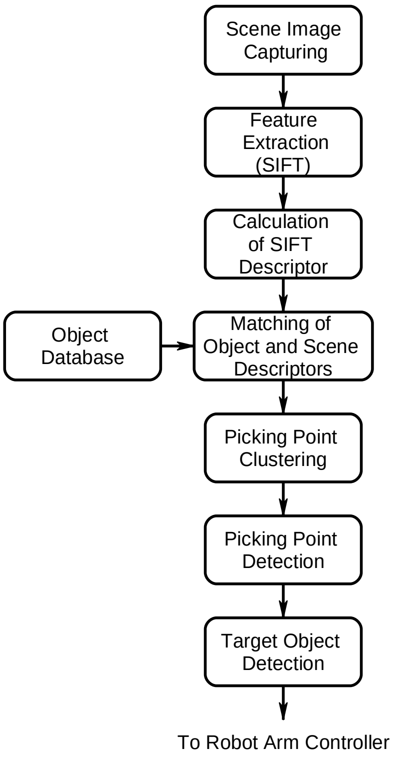

Fig.4: The on-line feature extraction on the scene and matching between the object and the scene

For the derived set M, the suitable and simplest approach for evaluating the registration transform between model(M)and image (I) is to estimate the planar homography using a least squares approach.

In Object localization phase, the clustering of possible grasping points is done and then the coordinates of the best grasping point is estimated.[1]

IV. Conclusions

In this paper, several object recognition and localization methods for machine vision have been studied along with the various challenges need to be met in real time applications such as clutter, scale changes, occlusion, illumination change, operating speed, etc. It has been found that some methods cope up with few challenges but at the same time they have proved inefficient while dealing with other. The SIFT technique is emerged the best one overcoming most of the challenges stated above. Thus, the SIFT method is going to be used in the proposed approach as a powerful tool for the enhancing the capability of machine vision to develop an autonomous and robust pick and place robotic system.

References

[1] P. Piccinini, A. Prati, R. Cucchiara, "Real-time object detection and localization with sift-based clustering", in proc. of Image and Vision Computing, 2012, pp. 1-15.

[2] K. Rahardja, A. Kosaka, "Vision-based bin picking: recognition and localization of multiple complex objects using simple visual cues", in IEEE/RSJ International Conference on Intelligent Robots And Systems, IEEE press, 1996, pp. 1448–1457.

[3] T. Knoll, R. Jain, "Recognizing partially visible objects using feature indexed hypotheses", IEEE J. Robot. Autom.2, 1986, pp. 3 –13.

[4] T. Gevers, A. Smeulders, "Color based object recognition", in Proc. of IEEE Int. Conference on Image Analysis and Processing, 1997, pp. 319 –326.

[5] M. Mueller, K. Segl, H. Kaufmann, "Edge- and region-based segmentation technique for the extraction of large, man-made objects in high-resolution satellite imagery", in Pattern Recognition, 2004, pp. 1619 –1628.

[6] A. Mian, M. Bennamoun, R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes", IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1584– 1601.

[7] A. Agrawal, Y. Sun, J. Barnwell, R. Raskar, "Vision guided robot system for picking objects by casting shadows", in Int. Journal of Robotics Research, 2009, pp. 1-28

[8] C. Cerrada, S. Salamanca, A. Adán, E. Pérez, J. Cerrada, and I. Abad, "Improved Method for Object Recognition in Complex Scenes by Fusioning 3-D Information and RFID Technology" In IEEE Transactions On Instrumentation And Measurement, 2009, pp. 3473–3480.

[9] Y. Yang, Q. Cao "Monocular vision based 6D object localization for service robot's intelligent grasping" In International Journal of Computers and Mathematics with Applications, 2012, pp. 1235–1241.

[10] K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors", in IEEE Trans. Pattern Anal. Mach. Intell., 2005, pp.1615 –1630.

[11] R. Jukjeon, G. Suji,S. Yongin, D. Gyeonggi, Journal of Internet Services and Information Security (JISIS), August 2015, pp. 48-57