Proceedings of the 2014 International Conference on
Advanced Mechatronic Systems, Kumamoto, Japan, August 10-12, 2014

Robust Vision-Based Detection and Grasping Object for Manipulator using SIFT Keypoint Detector

Widodo Budiharto
School of Computer Science
Bina Nusantara University
Jakarta- Indonesia
Email:wbudiharto@binus.edu

Abstract— The ability for a manipulator to detect and grasp an object accurately and fast is very important. Vision-based manipulator using stereo vision is proposed in this paper in order able to detect and grasp an object in a good manner. We propose a framework, fast algorithm for object detection using SIFT(Scale Invariant Features Transform) keypoint detector and FLANN (Fast Library for Approximate Nearest Neighbor) based matcher. Stereo vision is used in order the system knows the position (pose estimation) of the object. Bayesian filtering implemented in order to reduce noise from camera and robust tracking. Experimental result presented and we analyze the result.

Keywords: manipulator, stereo vision, SIFT Keypoint, FLANN, matching, Bayesian filter.

I. INTRODUCTION

One of the most common of manipulation tasks is grasping an object. Tasks performed by humans involve some form of grasping action. In the absence of feedback, the grasping action cannot be completed effectively. A human being grasps an object almost invariably with the aid of vision. We use visual information to identify and locate the object, and then decide how to grasp them.

Most work in robotic manipulation assumes a known 3-D model of the object and the environment, and focuses on designing control and planning methods to achieve a successful and stable grasp in simulation environments. Vision-based grasping is usually preceded by a number of tasks that effect the final grasping action. Based on the previous literature (visual-servoing) is huge and largely unorganized. In image based visual servoing, 2D image measurements are used directly to estimate the desired movement of the robot. Typical tasks like tracking and positioning are performed by reducing the image distance error between a set of current and desired image features in the image plane.

In manipulator, a link considered as a rigid body defining the relationship between two neighbouring joint axes. A link can be specified by two numbers, the link length and link twist, which define the relative location of the two axes in space. Joints may be described by two parameters. The link offset is the distance from one link to the next along the axis of the joint. The joint angle is the rotation of one link with respect to the next about the joint axis.

A variety of methods have been proposed to solve vision-based manipulation [1-5]. They use vision to aid just one of the above mentioned steps. In the past, most approaches to robotic grasping [6][7] assume availability of a complete 3-D model of the object to be grasped. In practice, however, such a model is often not available, beside the 3D models obtained from a stereo system are often noisy with many points missing, and 3-D models obtained from a laser system are very sparse. This makes grasping a hard problem in practice. In more general grasping, Kamon et al. [8] used Q-learning to control the arm to reach towards a spherical object to grasp it using a parallel plate gripper. Edsinger and Kemp [9] grasped cylindrical objects using a power grasp by using visual servoing and do not apply to grasping for general shapes. An inverse kinematics solver is proposed in [10] to find all joint angles for given position of the effectors on the manipulator. The target object is recognized by color segmentation. The 3D position is computed by the stereo vision system after contour extraction.

Inverse kinematics of manipulator and object location are the key technology for arm robot. We study various visual feedback methods from previous literature and develop a new model for grasping a bottle. We know that Extraction of image information and control of a robot are two separate tasks where at first image processing is performed followed by the generation of a control sequence. A typical example is to recognize the object to be manipulated by matching image features to a model of the object and compute its pose relative to the camera (robot) coordinate system.

The main contribution of this paper is a framework with low cost stereo camera and object detector using SIFT keypoint detector that very fast and suitable for manipulator in education. In this paper we develop algorithms for robust object detector and grasping system for manipulator or arm robot. First the target object is recognized by the vision system which than estimates the 3D pose of the object. Based on this information, the controller coordinates to move the arm robot to grasp the object/bottle. The framework proposed in this experiment shown in fig. 1 below, for example, the stereo camera for pose estimation attached about 50cm at the side of manipulator.

978-1-4799-6381-2/14/$31.00

Figure 1. Proposed framework for vision-based grasping using Computer Vision and stereo camera at the side of the manipulator.

II. VISION-BASED GRASPING MANIPULATOR

A. DOF Manipulator

We developed a framework of vision-based arm robot using 4 DOF (Degree of Freedom) arm robot from Lyxmotion that able to delivers fast, accurate, and repeatable movement. The forward kinematics process consists of computing the position and orientation of a robot end-effector pose resulting from a set of joint variables values, the robot manipulator Denavit-Hartenberg (DH) parameters are needed. When analyzing the kinematics of a robot, the rist step is to produce the forward kinematics equations that relate the desired end-effector pose matrix to the unknown joint values (q_1,q_2,…,q_n). For an n-joint robot, the forward kinematics equation in matrix form is :

$$\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_{3,\ldots,} \mathbf{A}_n = \mathbf{P} \qquad (1)$$

Where $\mathbf{A}_i$=$\mathbf{A}_i(q_i)$, a function of joint variable q_i, s the homogeneous matrix for link-frame F_i to link frame F_{i-1} and **P** is the end-effector pose[11]. To analyze the motion of robot manipulators, reference frames are attached to each link starting from frame $\mathbf{F}_0$, attached to the fixed link, all the way to frame $\mathbf{F}_n$, attached to the robot end-effector (assuming the robot has n joints). The DH model is obtained by describing each link frame along the robotic chain with respect to the preceding link frame. The DH parameters, routinely noted as d_i, a_i, α_i and Θ_i, relate geometrically to the frame location as shown in figure 2:

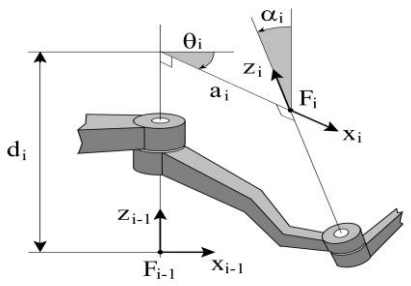

Figure 2. general DH Parameters for manipulator

The robot features used in this research are: base rotation, single plane shoulder, elbow, wrist motion, a functional gripper, and optional wrist rotate as shown in figure 3. In a manipulator system, given the angles of the joints, the kinematics equations give the location of the tip of the arm. Inverse kinematics refers to the reverse process, given a desired location for the tip of the manipulator, what should the angles of the joints be so as to locate the tip of the arm at the desired location. This manipulator is an affordable system with a time tested rock solid design that will last and last.

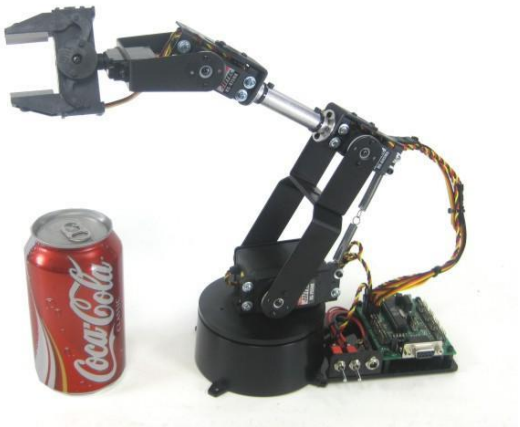

Figure 3. 4 DOF arm robot using stereo vision used in the experiment

The specification of this manipulator:

Base: Height = 6.9 cm
Hand/Grip: Max Length (No Wrist Rotate) = 8.7 cm
Hand/Grip: Max Length (With LW Rotate) = 11.3 cm
Hand/Grip: Max Length (With HD Rotate) = 10.0 cm
Length: Forearm = 12.7 cm
Length: Arm = 11.9 cm

We have developed a system for depth estimation to measure distance of object using low cost stereo camera Minoru 3D. We use OpenCV 2.4.8[12] for image processing. OpenCV software and a color filter algorithm are used to extract the specific color features of the object. Then, the 3D coordinates of the object to be grasped are derived by the stereo vision algorithm, and the coordinates are used to guide

the manipulator to the approximate location of the object using inverse kinematics.

B. *Object Detection and Grasping Model*

In our previous work [13], we detect an object based on the color; because of its drawback, now we improve the object detection system by giving a template of object for training. An object detection system has been developed that uses a new class of local image features. The features are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection. Features are efficiently detected through a staged filtering approach that identifies stable points in scale space [14]. The first stage of keypoint detection is to identify locations and scales assigned under differing views of the same object. Detecting locations that are invariant to scale change of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as scale space. It has been shown by Koenderink [15] and Lindeberg [16] that under a variety of reasonable assumptions the only possible scale-space kernel is the Gaussian function. Therefore, the scale space of an image is defined as a function, $L(x, y, \sigma)$, that is produced from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with an input image, $I(x, y)$

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

Where * is the convolution operation in x and y, and:

$$\mathbf{G(x, y, \sigma) = (1/2\pi\sigma^2) e^{-(x^2+y^2)/2\sigma^2}} \quad (3)$$

To efficiently detect stable keypoint locations in scale space, David G. Low proposed using scale-space extrema in the difference-of-Gaussian function convolved with the image

$$D(x, y, \sigma) = [G(x, y, k\sigma) - G(x, y, \sigma)] I(x, y)$$
$$= L(x, yk\sigma) - L((x, y, \sigma)) \quad (4)$$

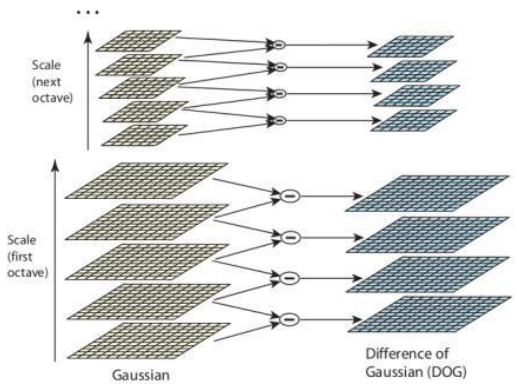

Figure 4. Gaussian scale-space pyramid create an interval in the difference-of-Gaussian pyramid.

Laplacian of Gaussian acts as a blob detector which detects blobs in various sizes due to change in σ. In short, σ acts as a scaling parameter. We can find the local maxima across the scale and space which gives us a list of (x,y,σ) values which means there is a potential keypoint at (x,y) at σ scale. But this LoG is a little costly, so SIFT algorithm uses Difference of Gaussians which is an approximation of LoG. Difference of Gaussian is obtained as the difference of Gaussian blurring of an image with two different σ, let it be σ and kσ. This process is done for different octaves of the image in Gaussian Pyramid. It is represented in below image. Once this DoG are found, images are searched for local extrema over scale and space. In order to detect the local maxima and minima of G(x, y, σ), each sample point is compared to its eight neighbors in the current image and nine neighbors in the scale above and below:

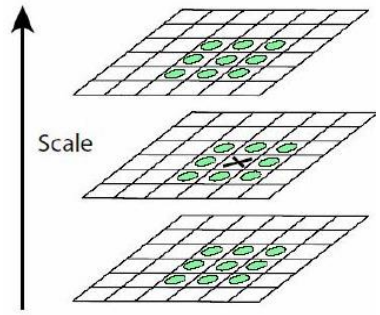

Figure 5. Maxima and minima detection in the difference-of-Gaussian image.

Lowe [17] proposed how to extracting keypoints and computing descriptors using the Scale Invariant Feature Transform (SIFT). Keypoints are detected using scale-space extrema in difference-of-Gaussian function D and efficient to compute. Grasp determination is probably one of the most important questions from manipulation point of view. Usually, the object is of unknown shape. The FLANN based matcher is object detector to make this system robust compared than color-based detector, we improve the FLANN algorithm to create rectangle line based on the average key points of the object as shown in figure 6 below:

Figure 6. Robust Object detector using FLANN based matcher for manipulator, rectangle line used to get center position of the object.

C. *Bayesian Filter*

Camera as vision sensor sometimes have distortion, so Bayesian decision theory used to state estimation and determine the optimal response for the robot based on inaccurate sensor data. We measure the probability of the absence of the object based on the calculation of Bayesian filter. If x is a quantity that we would like to infer from y, the probability $P(x)$ will be referred to as prior probability distribution. The Bayesian update formula is applied to determine the new posterior $P(x|y,z)$ whenever a new observation vector z is obtained [18]:

$$P(x|y,z) = P(y|x,z)P(x|z)/P(y|z) \quad (5)$$

Bayes filters represent the state at time t by random variables x_t . At each point in time, a probability distribution over x_t, called belief, $\mathrm{Bel}(x_t)$ represents the uncertainty. Bayes filters aim to sequentially estimate such beliefs over the state space conditioned on all information contained in the sensor data. To illustrate, let's assume that the sensor data consists of a sequence of time-indexed sensor observations $z_1, z_2, \ldots, z_t$. The belief $\mathrm{Bel}(x_t)$ is then defined by the posterior density over the random variable x_t conditioned on all sensor data available at time t:

$$Bel(x_t) = P(X_t|z_1, z_2, z_3, \ldots, z_t) \quad (6)$$

If η is normalization, then the formula becomes:

$$Bel(x_t) = \eta P(z_t|x_t) \int P(x_t|u_{t-1}, x_{t-1}) Bel(x_{t-1}) dx_{t-1} \quad (7)$$

Bayesian decision rule probabilistically estimate a dynamic system state from noisy observations. To improve the robustness of the vision system to track the object, we use Bayesian filter in continuous case, where probability labeled as P, action u, the outcome state x and x' as previous state. Integrating the outcome of actions in continuous case:

$$P(x\}u) = \int P(x|u,x')P(x')dx' \quad (8)$$

The proposed algorithms for detect an object/bottle, grasp it and move to the destination are shown below, to make sure that the object is really detected and tracking successfully using Bayesian filtering. Algorithm 1 with the main function named **DetectandCreateRectangle()** shows how to detect and get center coordinate X and Y for an object. Algorithm 2 is general algorithm for robust vision-based grasping for manipulator using SIFT keypoint detector.

Algorithm 1: Detection and creating rectangle line of the object

```
Function DetectandCreateRectangle()
begin
detect the object/bottle using SIFT Keypoint
if (detected & good_matches.size()>=4)
        summation of  the coordinate  X and Y position of the object
endif
//calculate average coordiate of x and y of the object
Average_X=average_X/good_matches.size()
Average_Y=average_Ygood_matches.size()
create rectangle line of the object
end
```

Algorithm 2: Robust Vision-based grasping for manipulator

```
do
call DetectandFindRectangle()
initialize bayesian  filter's variables
if object/bottle detected then
   begin
    update bayesian data
    move  arm robot to the object/ bottle
    move the gripper to the center of the object/ bottle
    if the position equal
    grasp the object/ bottle
    else
    move to gripper to the center of the object/ bottle
   end
endif
move the object to destination
go to the initial position
loop
```

III. EXPERIMENTAL RESULT

This paper has studied the vision-based arm robot with the goal of this project is to build a general purpose household robot. To determine the state of the object, the "good" grasp position should be first determined. The experiment conducted at our lab to grasp and move a bottle to the destination. We use Lynxmotion RIOS (Robotic Arm Interactive Operating System) SSC-32 software to configure and control the manipulator. To calculate and test inverse kinematics, we use general formula in Ms. Excel provided from the factory of the manipulator shown in figure 7.

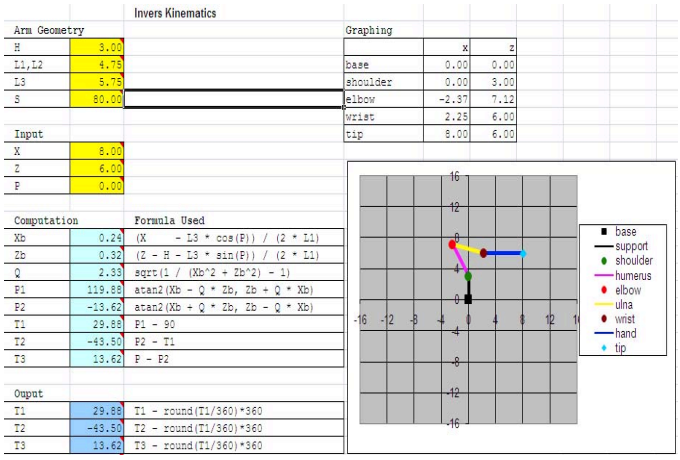

Figure 7. inverse kinematics for manipulator[19]

Program for object detection successfully developed with OpenCV with FLANN based matcher with average detection time is 33 ms and the manipulator is able to grasp it to move to other position. The FLANN based matcher is object detector to make this system robust compared than color-based detector as shown in figure 8.

Figure 8. Object detection and grasping by manipulator. Distance of the object also obtained using depth estimation from stereo camera.

After configuring and calibrating the servos of arm robot. We put an object/bottle in front of the manipulator to be grasped and move to other position. Based on our experiment we get the expected result as shown in table 1.

TABLE I. RESULT OF DETECTING AND GRASPING AN OBJECT IN 20X

No	Action	With Bayesian filter	Without Bayesian filter
1	Identify the object as bottle	100%	90%
2	Tracking an object without error	95%	80%
3	Grasping an object correctly	90%	80%
4	Estimate the distance of the object	90%	90%

a.

b. Based on table 1, shows that with the action identify the object as cup, percentage for success is 100% using Bayesian filter. With the action grasping an object correctly and estimate the distance of the bottle, percentage for success is 90% and percentage for failure is 10%. With Bayesian filter, the percentage of the system to track an object is higher. We propose robust system using SIFT keypoint detector to detect and object, and FLANN based matcher suitable to be used for real situation The accuracy and robustness of the system and the algorithm were tested and proven to be effective in real scenarios.

REFERENCES

[1] A. Bendiksen and G. D. Hager. “A vision-based grasping system for unfamiliar planar objects“. In ICRA, pages2844–2849, 1994.

[2] H. I. Christensen and P. L. Corke. “Visual servoing“. I. J.Robotic Res., 22(10-11):779–780,2003.

[3] D. Kragic and H. I. Christensen. “A framework for visual servoing“. In ICVS, pages 345–354, 2003

[4] A. Bicchi and V. Kumar, “Robotic grasping and contact: a review,” in ICRA, 2000.

[5] J. Jiao, Embedded Vision-Based Autonomous Move-to-Grasp Approach for a Mobile Manipulator,International Journal of Advanced Robotics System, vol. 9, pp.1-6, 2012.

[6] K. Shimoga, “Robot grasp synthesis: a survey,” IJRR, vol. 15, pp. 230–266, 1996.

[7] D. Purwanto. Visual Feedback Control in Multi-Degrees-of-Freedom Motion System. PhD thesis at Graduate School of Science and Technology - Keio University, Japan, 2001.

[8] I., Kamon, T. Flash, and S. Edelman, “Learning to grasp using visual information,” in ICRA, 1996

[9] A. Edsinger and C. C. Kemp, “Manipulation in human environments,” in IEEE/RAS Int’l Conf on Humanoid Robotics (Humanoids06), 2006.

[10] Y. Yang, “Binocular Stereo Vision based Humanoid Manipulator Control”, 30th International Conference on Control, pp. 3996 - 4001, China. 2011.

[11] R. Manseur, “Robot Modeling and Kinematics”, Da Vinci Engineering Press, 2006.

[12] OpenCV.org

[13] W. Budiharto et. al, “The Framework of Vision -Based Grasping for Manipulator”, Annual Conference on Engineering and Information Technology, 28-30 March 2014, Tokyo.

[14] D.G. Lowe, "Object recognition from local scale-invariant features," International Conference on Computer Vision, Corfu,Greece (September 1999), pp. 1150-1157.

[15] J.J., Koenderink, 1984, “The structure of images”, Biological Cybernetics, 50:363-396.

[16] T. Lindeberg, 1993, “Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention”, International Journal of Computer Vision, 11(3): 283-318.

[17] D.G Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, 60, 2, pp. 91-110, 2004.

[18] S. Thrun, Probabilistic robotics, The MI Press (2006).

[19] Lynxmotion.com