Stereo Sensors-based Object Segmentation and Location for a Bin Picking Adept SCARA Robot

HEMA C.R., PAULRAJ M.P., NAGARAJAN R., SAZALI YAACOB

School of Mechatronic Engineering, Universiti Malaysia Perlis, Kangar, Perlis, Malaysia
hema@unimap.edu.my

ABSTRACT

In this paper we present a stereo vision system for segmentation of partially occluded objects and computation of object grasping point in bin picking environments. The stereo vision system was interfaced with an Adept SCARA Robot to perform bin picking operations. Most researches on bin picking involve combination of vision and force sensors, however in this research an attempt is made to develop a bin picking system using only vision sensors for bin pick and place operation. An algorithm to segment partially occluded objects is proposed. The proposed stereo vision system was found to be effective for partially occluded objects and in the absence of albedo effects. The results are validated through real time bin picking experiments on the Adept Robot.

INTRODUCTION

A Bin Picking Robot requires information of the object to be picked and its exact location with respect to the bin area. It is generally assumed that the topmost object will be desired object to be picked from a bin with scattered or piled objects. The stereo vision system proposed considers two aspects, one is the segmentation of the bin image to identify the topmost object and the other is the location of the grasping point *(x, y and z co-ordinates)* of the object with respect to the bin area. In bin picking environments object occlusions in the bin image pose a challenge to the segmentation process. Several papers have been published on bin picking algorithms, of which very few have considered occluded objects. The concept of bin picking in robots has been under research for almost two decades. A segmentation algorithm combining the edge detection and region growing techniques to determine the shape of an object is presented by Al-Hujazi *et al* [1]. The images are obtained from range sensors; the segmented image is used to give the hold site information to the gripper. This algorithm however, deals with individual objects only, occluded objects are not considered in this research. A fuzzy based approach to determine the hold sites for randomly placed similar objects in a bin environment has been developed by Tae Won Kim and Hong Suh [8]. Wherein design techniques are investigated to obtain membership functions for the darkness and the brightness to find hold sites with the vertical and horizontal fuzzy filters as against matched filters. Sarah Wang et al. [6] have reported a 3D structured light vision system with a 2D binary vision for bin picking of twisted tubular parts. The system uses a structured light based 3D vision for localizing an entire tube in the bin. If it does not succeed in finding the entire tube, it locates graspable fragments of the tubes. The picked tubes are placed on a backlit table for comparison with its stable models for pose determination and recognition. Rahardja and Kosaka [5] have presented an algorithm that identifies complex industrial parts and estimates their pose. Stereo images are used to match the views in estimating the pose. Landmark features extracted by segmentation are used for recognition.

Martin Berger *et al.*, [4] have presented a bin picking system with three independent vision guiding modules. A vacuum gripper picks an object using stereo sensors and structured light without any knowledge of the object. In the next module the pose of the object is compared with a CAD model of the object and object is mounted.The third module ensures correct mounting.This research also considers only non occluded objects randomly placed in a bin. Most researches on bin picking use vision only for object recognition and pose determination, while others use a model based approach which compares the object image with a model database for pose determination.

The stereo vision system proposed in this paper comprises of two machine vision cameras in a stereo rig.PULNIX progressive scan cameras were used in this experiment.A base length of 70mm was found to be suitable for the current application.A vacuum gripper was used for grasping the objects. Similar geometrical shaped objects with partial occlusions are used in the experimental process.The segmentation algorithms use binary thresholding techniques and image histogram to identify the topmost object in the bin as detailed in Section II. Section III elaborates on the object feature extraction process, while Section IV and V describe the neural network architecture and experimental results respectively. Section VI comprises the observations, conclusion and future research aspects of the paper.

SEGMENTATION

Image Acquisition and Preprocessing

The stereo vision system consists of two PULNIX TM 6702 machine vision cameras in a stereo rig for capturing the bin images. Direct lighting of the bin is avoided to reduce brightness and albedo effects. Objects in the bin partially occlude each other and have different intensity levels due to their location. The captured images are of size 640 x 480 pixels and are monochromatic.The acquired images are first pre-processed to improve the efficiency of the segmentation process.The pre-processing involves two stages i) image resizing and ii) filtering.To minimize the processing time and to improve the efficiency of the system without significant loss of information of the objects, the images are resized. A image size of 128 x 96 pixels was found to be suitable for the current application. Since the topmost object will be the one without occlusions and the one with higher intensity levels in comparison to the rest of the objects, filtering techniques are applied to smooth out the intensity of the object and to enhance its edge. The images are filtered using a regional filter and a mask. This masked filter filters the data in the image with the 2-D linear Gaussian filter and a mask the same size as the original image for filtering.This filter returns an image that consists of filtered values for pixels in locations where the mask contains ones and unfiltered values for pixels in locations where the mask contains zeroes. The above process smoothens the intensity of the image around the objects.The resulting filtered image is then subjected to segmenting techniques as detailed in the following section.

Bin Image Segmentation

Bin image segmentation involves identifying the top most object from the cluster of objects in the bin for pick up. Since all the objects are partially occluded except the topmost object, separating the topmost object can be done using the grey value of

the object. A histogram of the bin stereo images displays the grey levels of the image. Segmentation using binary thresholding is possible by identifying the pixels of grey levels higher than a threshold value which is assumed to relate to the topmost object, as the topmost object has a brightness level higher than the other objects in the bin. A suitable threshold segmenting only the topmost object is to be computed. For real-time bin picking, automatic determination of threshold value is an essential criterion. To determine this threshold value an algorithm is proposed which uses the grey levels of the image from the histogram of both the stereo images to compute the threshold. The proposed algorithm is as follows :

Step 1: The histogram is computed from the left and right gray scale images for a bin value of 0 to 255.

Counts a (i), i=1, 2, 3,..., 256

contains the number of pixels with a gray scale value of (i-1) pixels for the left image.

Counts b (i), i=1, 2,3,...,256

contains the number of pixels with a gray scale of (i-1) for the right image.

Step 2: Compute the logarithmic weighted gray scale value of the left and right image as

$$ta(i) = log(count\ a(i)) * (i-1) \quad (1)$$

$$tb(i) = log(count\ b(i)) * (i-1) \quad (2)$$

where $i = 1,2,3,\ldots,255$

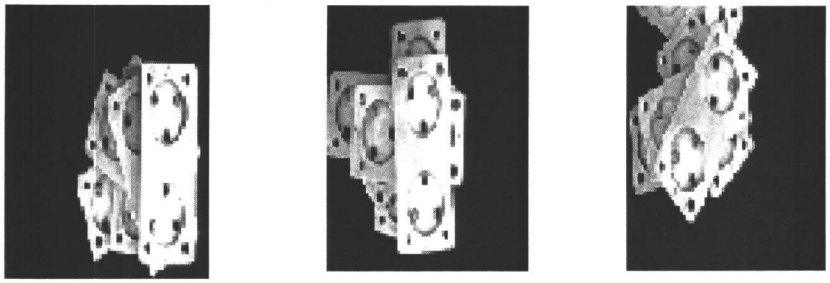
Step 3: Compute the logarithmic weighted gray scale

$$tam = \frac{1}{256}\sum_{i=1}^{256} ta(i) \quad (3)$$

$$tbm = \frac{1}{256}\sum_{i=1}^{256} tb(i) \quad (4)$$

Step 4: The threshold T is the maximum value of 'tam' and 'tbm'.

Figure 1. Shows the bin images before and after segmentation.



(a) Bin images

(b) Segmented images showing the top most object

Threshold of both the stereo images are computed separately and the max value of the two thresholds is applied as the threshold to both the images. The grey images are converted to binary images by applying the threshold derived from the histogram.

OBJECT FEATURE EXTRACTION

The next phase of the research is to compute the location of the object with respect to the bin area. To pick up the topmost object the robot is to be provided with the location co-ordinates of the object viz, x, y and z co-ordinates with respect to the bin area. The x and y co-ordinates can be determined from the 2D segmented image of the object, whereas the z co-ordinate which is the depth or distance co-ordinate requires the 3D representation. To find the z co-ordinate we propose a neural network and stereo image approach which trains a neural network to compute the distance or z co-ordinate of the object from its stereo images. The object features of the stereo images are used as the input data and the distance of the object from gripper is used as the output data to train the neural network. The object features are extracted using singular value decomposition. The following sections elaborate on SVD and feature extraction process.

Singular Value Decomposition

The Singular Value Decomposition (SVD) is a widely used technique to decompose a matrix into several component matrices, exposing many of the useful and interesting properties of the original matrix [9]. Any 'm x n' matrix **A** (m >= n) can be written as the product of a 'm x m' column-orthogonal matrix **U**, an 'm x n' diagonal matrix **W** with positive or zero elements, and the transpose of an 'n x n' orthogonal matrix **V** [6]:

$$A = U W V^t \tag{5}$$

where

$$W = \begin{bmatrix} w_1 & 0 & . & 0 & 0 \\ & w_1 & . & 0 & 0 \\ . & . & . & . & . \\ 0 & 0 & . & w_{n-1} & 0 \\ 0 & 0 & . & 0 & w_n \end{bmatrix} \tag{6}$$

and

$$U^t U = V V^t = I \tag{7}$$

where

$$w_1, w_2, \ldots, w_{n-1} w_n \geq 0$$

't' is the transpose and ***I*** is an identity matrix. The diagonal elements of matrix ***'W'*** are the singular values of matrix '**A**', which are non-negative numbers. The singular values obtained by the SVD of an image matrix give algebraic feature of an image, which represents the intrinsic attributes of an image [9].

Feature Extraction

To pick up the topmost object, the location of the object is to be computed. Since only one object is present in the segmented image, finding the centroid of the object will provide the x and y coordinates of the object location, however it is also important to know the z coordinate (distance) to move the gripper for pick up process. Stereo feature matching is one of the popular methods for distance computation [5]. In this research we present a method to compute the distance from the added stereo images of the object, called unified stereo imaging, wherein the stereo images of the object are added. The features of the added image varies with respect to the distance of the object, this aspect is used to train the neural network. To optimise the computational time edge images of the added image is used in the feature extraction process. Singular value features are extracted using SVD, for training the network singular value are extracted for various distance of the object. Figure 2 shows the flow diagram for object feature extraction. Singular values feature having values below '1' are ignored as few prominent singular value features of the object are sufficient to represent the object distance [2]. Hence ten significant features of the object image are fed to a neural network to train the network for object distance computation

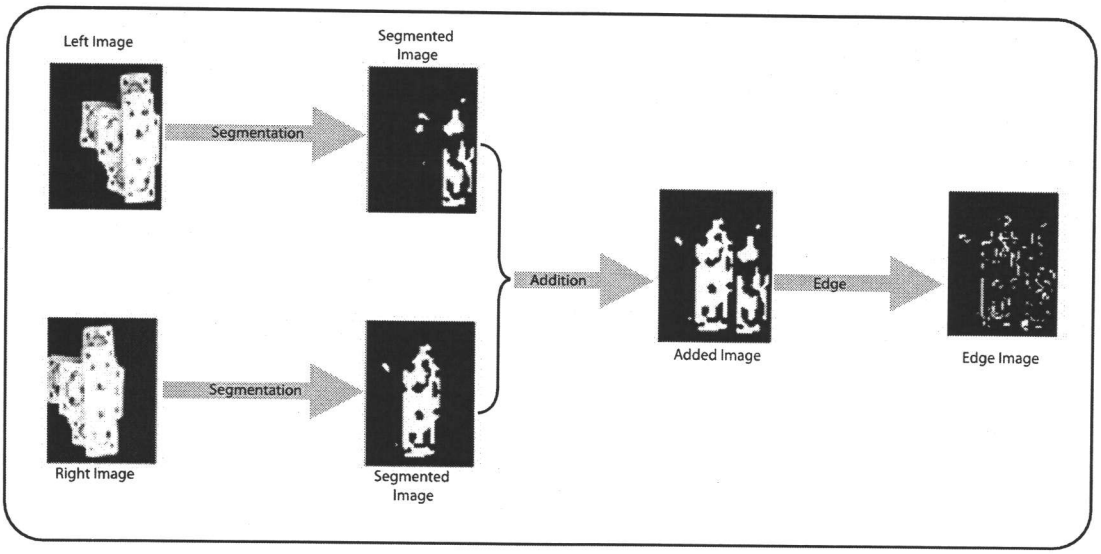


Figure 2. Flow Diagram for Object Feature Extraction.

NEURAL NETWORK ARCHITECTURE

The Neural Network architecture consists of three layers. 10 singular values are fed to the network as input data. The hidden layer is chosen to have 5 neurons and the output consists of 1 neuron, which represents the object distance. The hidden and input neurons have a bias value of 1.0 and are activated by bipolar sigmoid activation function. The initial weights for the network are randomized between -0.5 and 0.5 and normalized. The initial weights that are connected to any of the hidden or output neuron are normalized in such a way that the sum of the squared weight values connected to a neuron is one. This normalization is carried out using equation (8), which is used to implement the weight updating.

$$w_{1j}(new) = \frac{w_{1j}(old)}{\sqrt{w_{1j}^2 + w_{2j}^2 + \ldots + w_{nj}^2}} \quad ,j=1,2,3,\ldots,p \tag{8}$$

where n - number of input units and p - number of hidden units

A sum squared error criteria as defined by equation (9) is used as a stopping criteria while training the network. The sum-squared tolerance defined in equation (9) is fixed as 0.001. The network is trained by the conventional back propagation procedure.

Table 1. Training Parameters and Results of the Neural Network.

Network Parameters	Values
No. of input neurons	10
No. of Output neurons	1
No. of hidden neurons	5
Bias value	1.0
Learning Rate	0.1
Tolerance	0.001
Training time	70.7 seconds
Training Epoch	15112
Success Rate	92.45%

The cumulative error is the sum squared error for each epoch and is given by:-

$$\text{Sum squared error} = \sum_{p=1}^{p}\sum_{k=1}^{m}(t_k - y_k)^2 \qquad (9)$$

where

t_k is the expected output value for the k^{th} neuron,
y_k is the actual output value for the k^{th} neuron,
m is the total number of output neurons, and
p is the total number of input neurons.

EXPERIMENTAL RESULTS

In the training process 40 stereo images of the bin with similar objects are acquired. The bin images are pre-processed to smooth the intensity level of the object. The pre-processed images are segmented to extract the topmost object in the bin. The segmented left and right images are added and the edge of the added image is extracted. The 'x' and 'y' co-ordinates are computed from the left segmented object image, which is considered as the reference image. Finding the centroid of the reference image gives the 'x' and 'y' location co-ordinates. To compute the 'z' co-ordinate, the singular value features are extracted from the edge image using SVD. 10 singular values of an image are fed as input and the object distance is fed as output to a simple feed forward neural network. The network is trained by the back propagation training procedure.

In the testing phase the network is tested with 53 sample data. The proposed method is found to successfully compute the distance for 49 images with a success rate of 92.45% and error tolerance of 0.001. Figure 3. shows the Cumulative error versus epoch plot of the trained neural network. Table 1 shows the training parameters and test results of the Neural Network.

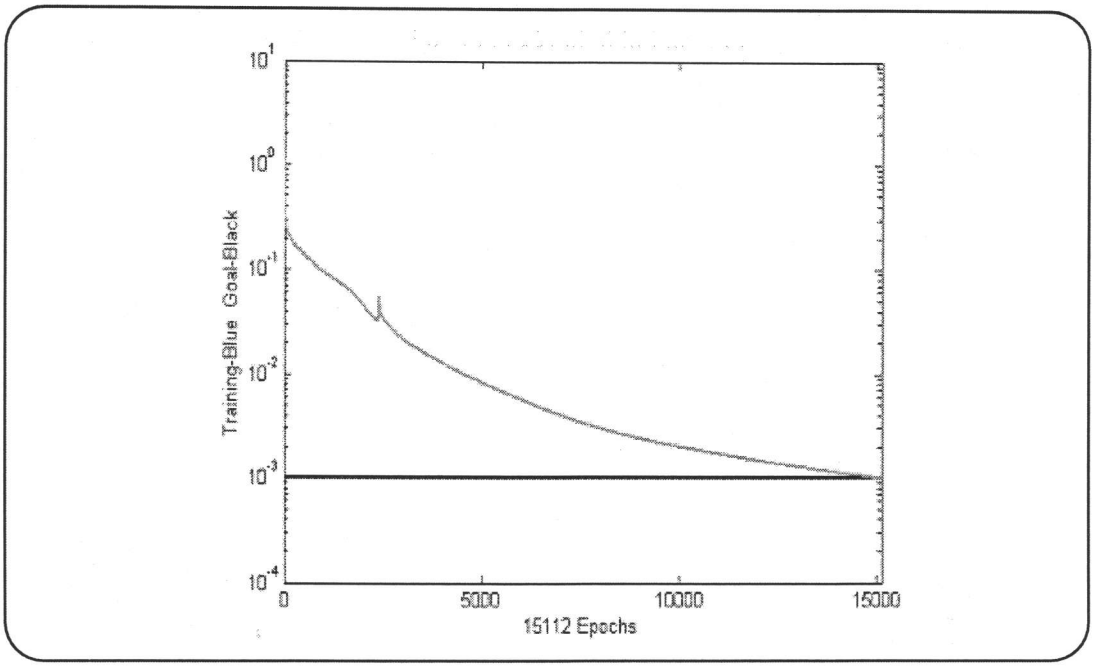


Figure 3. Cumulative errors versus Epoch Plot.

In the real time experimentation phase, the developed stereo vision system was interfaced with an Adept SCARA Robot shown in Figure 4. A vacuum gripper was used for pick and place operation. A bin with seven partially occluded objects was used for testing. The Adept SCARA Robot was tested for real time bin picking using the object grasping point ['x','y' and 'z' co-ordinates] computed by the stereo vision system. The bin picking system was successful in picking six out of the seven objects placed in the bin with an average bin pick and place performance of 85.7%

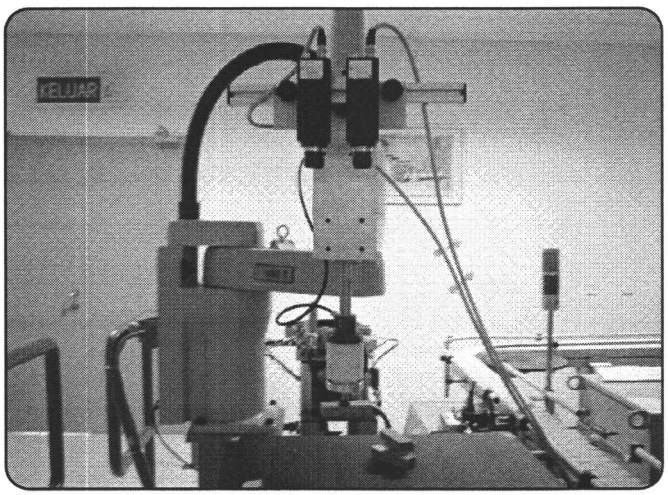


Figure 4. Adept Robot with stereo sensors and vacuum gripper.

CONCLUSION

A bin picking system using stereo vision sensors for object grasping point is presented. An algorithm for segmentation of partially occluded objects is presented. The system is experimentally verified and results presented are satisfactory. The major constraint of the proposed system was poor segmentation in the presence of albedo effects and uneven brightness in certain parts of the bin. Optimal lighting conditions are essential to derive satisfactory results. Future work will include optimising lighting conditions and improving the network performance. Test series proved the applicability of the proposed vision system in segmenting partially occluded objects and computing grasping points.

ACKNOWLEDGMENT

The authors thankfully acknowledge the funding [Code. No. 04-02-10-0024-PR001] received from IRPA Fund of the Malaysian Government.

REFERENCES

1. Ezzet Al-Hujazi and Arun Sood, R. (1990). Image Segmentation with Applications to Robot Bin-Picking Using Vacuum Gripper. *IEEE Transactions on System, Man and Cybernetics,* Vol. 20 (No.6), 1313- 1325.

2. Fausett, L. (1990). *Fundamentals of Neural Network Architecture and Applications.* USA: Prentice Hall.

3. Hong, Z.-Q. (1991). Algebraic Feature Extraction of Image for Recognition. *IEEE Transactions on Pattern Recognition,* vol. 24(No. 3), 211-219.

4. Martin Berger, G. B. a. S. S. (2000). Vision Guided Bin Picking and Mounting in a Flexible Assembly Cel. *Thirteenth International Conference on Industrial and Engineering Application of Artificial Intelligence and Expert Systems. U.S.A.*

5. Krisnawan Rahardja, A. K. (1996). Vision-Based Bin –Picking: Recognition and Localization of Multiple Complex Objects Using Simple Visual Cues' in Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems,* Vol: 3, 1448 - 1457.

6. Sarah Wang, R. C., Avi kak, Ichiro Kimura and Michiharu Osada. (1994). Model-Based Vision for Robotic Manipulation of Twisted tubular parts: Using affine transforms and heuristic Search. *IEEE Transactions* 208 - 215.

7. S.N.Sivanandam, M. P. (2003). *Introduction to Artificial Neural Networks.* India: Vikas Publishing House.

8. Suh, T. W. K. a. H. (1991). A Fuzzy Logic Based Bin Picking Technique. *IEEE IECON'91,* 1573 – 1578.

9. Watkins, D. S. (2002). *Fundamentals of Matrix Computations.* New York: Wiley Interscience Publications.