DE GRUYTER OPEN

1862
RIGA TECHNICAL UNIVERSITY

Electrical, Control and Communication Engineering

ISSN 2255-9159 (online)
ISSN 2255-9140 (print)
2017, vol. 12, pp. 34–39
doi: 10.1515/ecce-2017-0005
https://www.degruyter.com/view/j/ecce

An Object Detection and Pose Estimation Approach for Position Based Visual Servoing

Lei Shi (*MSc Student, Tallinn University of Technology*)

Abstract – **In this paper, an object recognition method and a pose estimation approach using stereo vision is presented. The proposed approach was used for position based visual servoing of a 6 DoF manipulator. The object detection and recognition method was designed with the purpose of increasing robustness. A RGB color-based object descriptor and an online correction method is proposed for object detection and recognition. Pose was estimated by using the depth information derived from stereo vision camera and an SVD based method. Transformation between the desired pose and object pose was calculated and later used for position based visual servoing. Experiments were carried out to verify the proposed approach for object recognition. The stereo camera was also tested to see whether the depth accuracy is adequate. The proposed object recognition method is invariant to scale, orientation and lighting condition which increases the level of robustness. The accuracy of stereo vision camera can reach 1 mm. The accuracy is adequate for tasks such as grasping and manipulation.**

Keywords – **Object recognition; Pose estimation; Stereo vision; Visual servoing.**

I. Introduction

Robotic manipulators can perform complicated grasp task fast and precisely with known environment. But there exists limitation. No feedback information for grasp task is provided. The trajectory and path are pre-programmed, if there is any change in the environment, for example, the object is missing, the robot system will not know. With visual perception, it brings vision feedback to the robot system to make it a close loop to control robot for grasping task. Such a vision based control is called visual servoing.

Generally, there are 3 types of visual servoing, image based (2D) visual servoing, position based (3D) visual servoing [1], [2] and 2 1/2 D visual servoing [3]. 2D visual servoing uses features from image for control and 3D visual servoing estimates the pose first and uses the pose information for control. 2 1/2 D visual servoing is a hybrid method which combines the 2D and 3D visual servoing. It decomposes the pose into position and orientation first and then controls them separately.

To accomplish the visual servoing, first step is object recognition. Secondly, the position and orientation of the object needs to be estimated. The third part is the controller design. In this paper only the first 2 steps are presented. The algorithm used for object detection is an edge segmentation method and the descriptor is based on RGB color information. A simple teaching stage is required in order to get the information of the object to be grasped. The algorithm is designed to have a descriptor invariant to the scale, orientation and light change so that the robustness is increased. Relevant studies have been done in [4], [5], where color co-occurrence is investigated for object recognition. An extensive evaluation on different color descriptors have been presented in [6]. To tackle the scale and orientation problem, SIFT [7] and SURF [8] descriptors are widely used. More recently, a descriptor called BRISK, which is even faster than SURF, is proposed in [9]. The descriptor proposed in this paper uses RGB color intensity and the standard deviation of RGB color distribution to correct color intensity threshold online. A simple teaching stage is required in order to get the information about the object. The algorithm is designed to have a descriptor invariant to the scale, orientation and light condition change within a certain range while the computational cost is relatively low.

Depth information is critical for pose estimation. A popular approach is using RGB-D camera like Kinect to obtain depth map. But the accuracy of Kinect is not suitable for robotic grasp. In [10], the depth accuracy is evaluated among other characteristics. Stereo vision camera can provide better depth accuracy. To estimate the pose, different approaches such as Extended Kalman Filter and Particle filter can be used. The solution used in this paper uses plane to represent the object. Then the problem is transferred to align the plane of current object in scene with the desired one. The plane is represented by 3 points which are not in a line. The 3 feature points representing the desired plane and object plane are selected based on the same rule. By aligning these 2 sets of points, 2 planes are also aligned. In order to align 2 sets of point, iterative closest point (ICP) and random sample consensus (RANSAC) could be used. The criterion is to minimize the accumulated error of transformation. The method used in this paper is based on [11]–[14]. Two feature point sets represent the target pose and object pose. The distance between these two sets of points is minimized so that a transformation matrix between the object pose and desired pose is obtained. From the transformation matrix, the pose can be easily derived for further use.

Although the idea of visual servoing has been brought up for a long time and the topic has been discussed a lot, still there are not a lot of successful practical integrations in the field. Robustness, accuracy and convenience are important factors that need to be considered. The proposed approach is trying to lower the requirement of the *a-priori* knowledge about the objects and scenes while different objects with different characteristics can be recognized and their poses can be estimated effectively. The proposed approach for object detection and recognition is tested in different environments with different variables for validation of robustness. Depth accuracy of stereo camera is also measured in an experiment in order to validate the accuracy for visual servoing tasks.

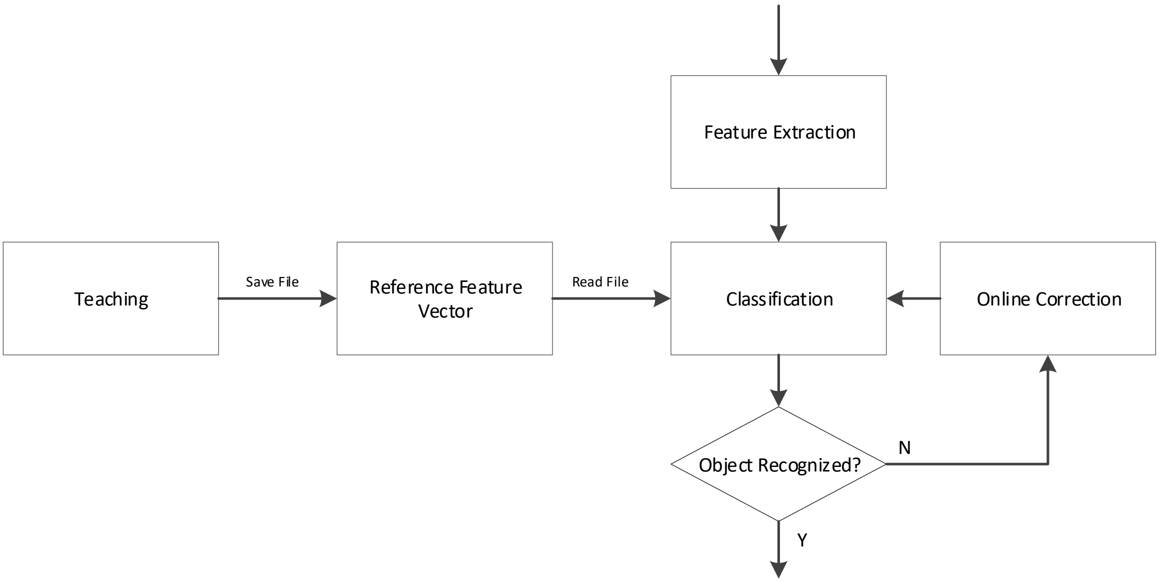

Fig. 1. Overview of object recognition.

The rest of this paper is organized as follows: in Section II, the proposed object detection and recognition algorithm is explained in detail; Section III explains how stereo camera is constructed and presents the pose estimation method; Section IV explains how the experiment is setup and conducted, as well as explains the methodology. Then the results are presented and evaluated. Finally a discussion of the results and future work are presented in Section V.

II. Object Recognition

The object detection algorithm needs to be robust enough. To be more specific, it should be robust in orientation and scale of descriptor and lighting condition. During the motion of robotic manipulator, the scale and orientation of object is changing. The lighting condition is also changing due to environment change. If camera loses the object during servoing, it can cause problems. Thus increasing the robustness of object detection can decrease the probability of losing the object.

The proposed algorithm uses edge-based segmentation for object detection and the online correction is based on color-based descriptor. A simple training stage is required for object detection and recognition, as shown in Fig.1. In this stage, the feature descriptor, color intensity and color distribution are recorded as reference. Later the reference feature vector is read for classification. Reference color intensity is calculated as in (1)

$$c_{\mathrm{r}} = \frac{1}{n}\sum_{i=1}^{n}(R_i + G_i + B_i), \quad (1)$$

where n is the size of the area of contour, R, G and B is the pixel value in red, green and blue plane of image, respectively.

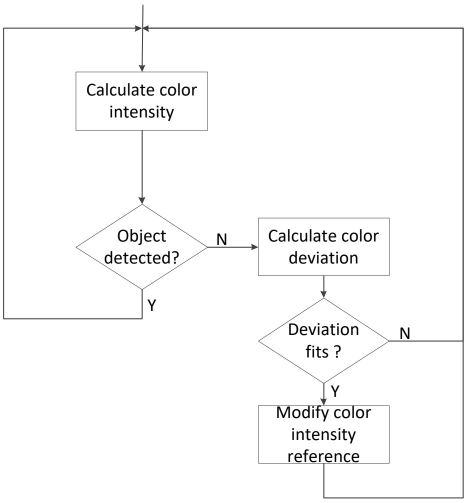
Calculate color intensity
Object detected?
N
Y
Calculate color deviation
Deviation fits ?
N
Y
Modify color intensity reference

Fig. 2. Online correction algorithm.

The descriptor is vulnerable to the light condition change as it is linear to the light intensity. Color intensity will be higher when the environment is brighter, and lower when it is darker.

When environment light changes, the classifier may not be effective and this means that it is variant to light condition change. The online correction algorithm is used to solve the problem. Figure 2 is the overview of the online correct algorithm for object recognition.

Reference color distribution is the standard deviation of pixel value in red, green and blue plane (2).

$$\sigma_{\mathrm{r}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)}, \quad (2)$$

where σ_{r} is ($\sigma_{\mathrm{r}}\ \sigma_{\mathrm{g}}\ \sigma_{\mathrm{b}}$), x is pixel value and $\overline{x}$ is mean pixel value.

The object colour intensity c_{o} is calculated. Whether object is detected or not is determined by (3).

$$I = \begin{cases} \text{True}, & \dfrac{|c_\mathrm{r} - c_\mathrm{o}|}{c_\mathrm{r}} < T_\mathrm{colour}, \\ \text{False, else} \end{cases} \tag{3}$$

where I is recognition result using colour intensity, T_colour is the threshold. If object is not detected, then color distribution is calculated, in (4) it explains whether the color distribution fits the reference value.

$$D = \begin{cases} \text{True}, & \dfrac{|\sigma_\mathrm{r} - \sigma_\mathrm{o}|}{\sigma_\mathrm{r}} < T_\mathrm{d}, \\ \text{False,else} \end{cases} \tag{4}$$

where T_d is threshold. If D is true, then c_o is assigned to c_r. and the correction is finished. If the color deviation is within the threshold, then reference color intensity is modified with the new value. And this new value is used as new color intensity reference.

III. Pose Estimation

Pose estimation is essential for position based visual servoing. The error between the desired pose P_r and object pose P_o is fed to visual controller. Both P_r and P_o are in camera coordinate frame. The depth is first calculated by using stereo vision technique, and then the transformation matrix T_c, which consists of rotation matrix R_c and translational vector $\boldsymbol{t_c}$, is obtained. In general, the pose is estimated by aligning 2 sets of points. 3 feature points are selected representing the desired plane and object plane. By aligning these 2 sets of points, 2 planes are also aligned.

A. Stereo Vision

Camera model [15] is pinhole camera model (5).

$$k \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = KE \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \tag{5}$$

where k is scaling factor, u and v are image coordinates, K is intrinsic camera matrix and E is extrinsic camera matrix. Proper calibration needs to be performed to ensure the accuracy of stereo vision system. A chessboard pattern is used for calibration. First, both left camera and right camera are calibrated separately. Then, the intrinsic and extrinsic camera matrices K_l, E_l and K_r, E_r are obtained. A stereo calibration is performed hereafter and the rotation matrix R_s and translational vector $\boldsymbol{t_\mathrm{s}}$ between 2 cameras are obtained. After calibration, the distortion coefficient is obtained and thus image distortion is also removed accordingly.

Rectification is performed before finding point correspondence. The purpose of rectification is narrowing down the search range of epipolar lines. In order to find point correspondence on epipolar line, it is necessary to search in both x and y direction in image coordinates. After rectification, the image planes are parallel and epipoles are moved to infinity, the epipolar lines are parallel as well so that the search is only required along one direction.

For solving the correspondence problem, blocking of matching algorithm is used. The final disparity map combines 2 disparity maps with different disparity range (6).

$$I_\mathrm{dtot}(u,v) = \begin{cases} I_1(u,v), & \text{if } I_1(u,v) \neq 0; \\ I_2(u,v), & \text{else,} \end{cases} \tag{6}$$

where $I_1(u,v)$ and $I_2(u,v)$ are disparity maps with different disparity range. Depth Z is then calculated in (7).

$$Z = \frac{bf}{disparity \cdot \mu}, \tag{7}$$

where b is baseline, f is focal length and μ is pixel size.

B. Description of Pose Estimation

Method (8)–(13) for pose estimation is derived from [14]. Feature point p_d is desired place points and p_o is object plane points. The general idea is to minimize the accumulated error of geometric distance between p_d and p_o (8).

$$\Sigma_\mathrm{e} = \sum_{i=1}^{3} \left\| R_c p_\mathrm{o} + t_c - p_\mathrm{d} \right\|. \tag{8}$$

Minimizing of accumulated error involves term R_c and $\boldsymbol{t_c}$. However, the accumulated error can be decomposed so that only R_c is involved. To solve decomposed accumulated error, SVD can be applied and this non-iterative method is tested and proven faster than ICP method. The centroid of p_d and p_o is calculated in (9).

$$p_\mathrm{dcen,ocen} = \frac{1}{3}\sum_{i=1}^{3} p_{\mathrm{d}i,\mathrm{o}i}. \tag{9}$$

The covariance matrix is constructed in (10) by using the sets of points and their centroids.

$$M = \sum_{i=1}^{3} (p_{\mathrm{o}i} - p_\mathrm{ocen})(p_{\mathrm{d}i} - p_\mathrm{dcen})^\mathrm{T}. \tag{10}$$

Then singular value decomposition (SVD) is performed on M (11). R_c is then calculated by unitary matrix U and V (12). In (13), $\boldsymbol{t_c}$ is calculated.

$$(U,S,V) = SVD(M); \tag{11}$$

$$R_c = VU^\mathrm{T}; \tag{12}$$

$$\boldsymbol{t_c} = -R_c p_\mathrm{ocen} + p_\mathrm{dcen}. \tag{13}$$

The transformation matrix is in camera coordinate frame. For later use, it will be transformed into TCP (Tool Center Point) coordinate frame.

IV. Results

The cameras used are two Leopard LI-USB30-M021C usb cameras. Pixel size of camera is 3.75 μm. Camera focal length is 6 mm and baseline is set to 200 mm. Cameras are placed in parallel configuration. In the first part of this section, the results of object recognition are presented and in the second part, accuracy of stereo vision camera is presented and analyzed.

A. Object Recognition

The object recognition algorithm is tested to verify the design purpose of the algorithm. In the experiment, the algorithm is tested with 3 different objects in an indoor environment. The first one is a tea box with a mark of black circle. The second one is a tea box with a mark of randomly selected colored image. The third one is the same colored image on a cup. Each object is tested in 4 different poses both in dark and bright lighting conditions. Dark lighting condition is achieved by turning off the lamps in the room, and the bright lighting condition is with lamps on. With this experiment setup, it can be tested verified whether the proposed descriptor is invariant to the orientation and scale in different lighting condition. The purpose of using markers is to eliminate the disturbance so that the validation of the color based descriptor and online correction algorithm could be verified easily. The results of the first experiment are displayed in Fig. 3 – Fig. 8.

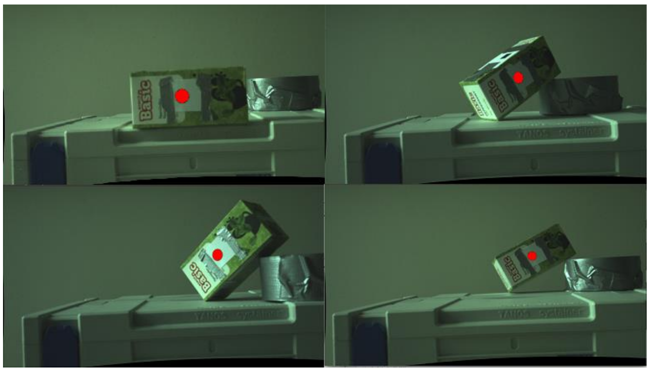

Fig. 3. Object recognition with black circle mark from different positions and orientation in dark condition.

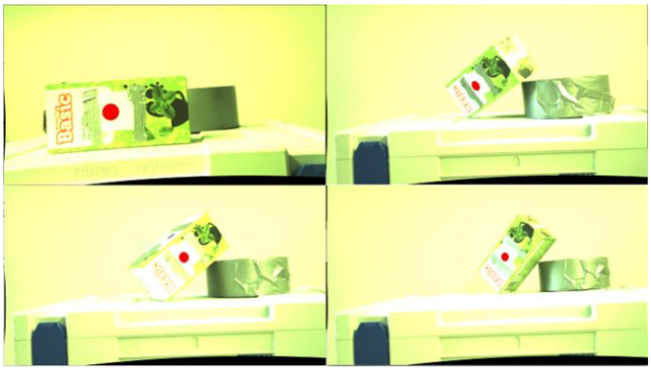

Fig. 4. Object recognition with black circle mark from different positions and orientation in a bright light condition.

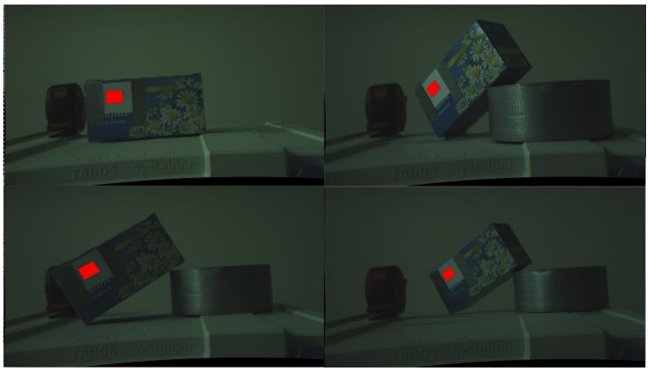

Fig. 5. Object recognition with a mark of colored image from different position and orientation in a dark condition.

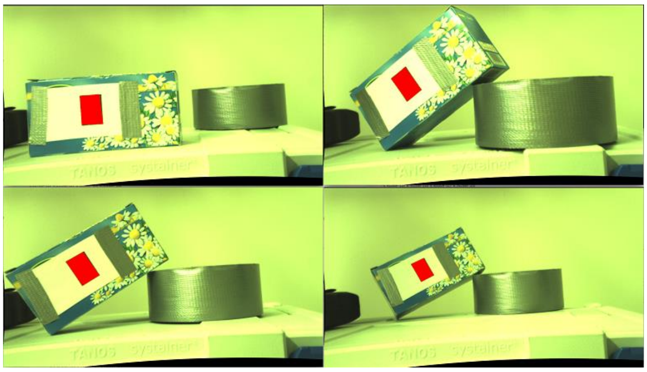

Fig. 6. Object recognition with a mark of colored image from different positions and orientation in a bright light condition.

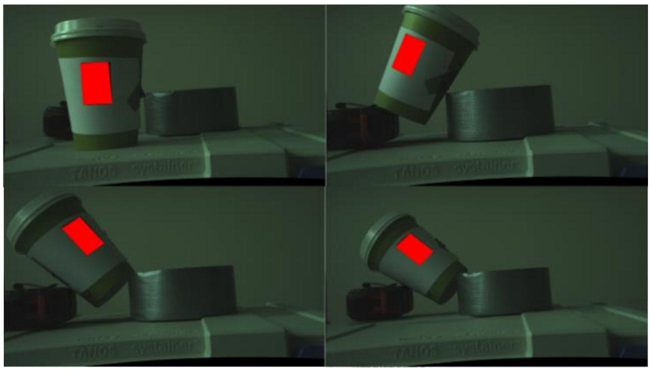

Fig. 7. Object recognition with a mark of colored image on a cup from different positions and orientation in dark condition.

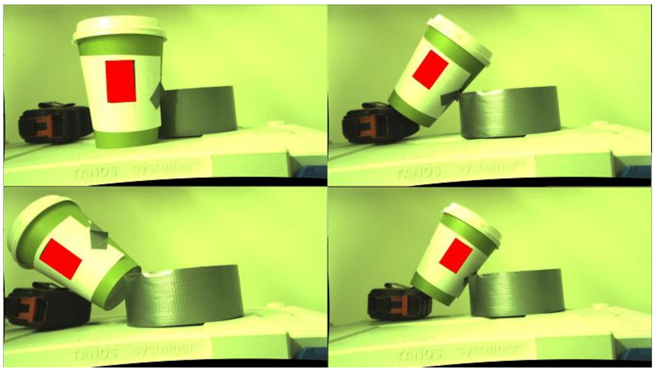

Fig. 8. Object recognition with a mark of colored image on a cup from different positions and orientation in a bright light condition.

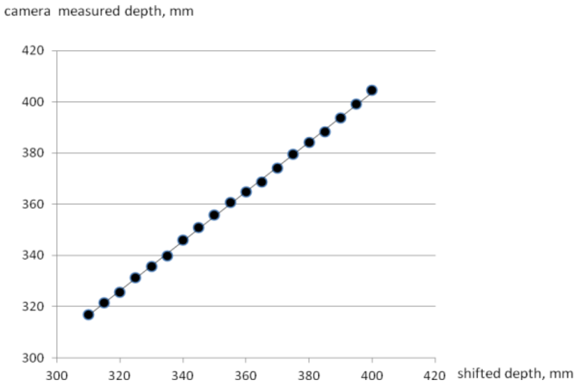

Fig. 9. Camera measured depth versus shifted depth.

As indicated in Fig. 3 to Fig. 8, the recognized contour is drawn in red color. All objects are recognized from different poses in both dark and bright lighting conditions. The 4 different poses tested cover different positions and orientation, which proves the invariance to scale and orientation. Using black mark and color pattern in the experiment, the proposed descriptor is shown effective on both textured or texture-less pattern. The different shapes, circle and rectangle, are also compared, which indicates that the descriptor is also invariant to shape. Putting the color pattern on a cup evaluates the validation of a non flat surface.

B. Stereo Vision

The number of images taken for calibration is 14. The calibration pattern used is a classic chessboard pattern. The chessboard pattern is placed in different poses while it is visible in both left and right images. The result of calibration is summarized in Table I. In single calibration, both the errors of left and right camera are less than 0.5 pixel. For stereo camera calibration, the error exceeds just 1 pixel.

To test the accuracy of the depth of stereo camera, the following experiment is carried out. The depth value is calculated with the following method. A small window of size 11 by 11 pixels is drawn on the image. The mean depth value within this small window is considered as camera measured depth value. In the experiment, 20 measures are conducted. Stereo camera is adjusted so that the small window whose center coincides with image center is on the same part of object in all the measures. In each measure, the object is placed 5 mm away from the cameras. The depth information is measured and analyzed. However it is difficult to measure the absolute depth from the object to camera center. The structure and dimension within camera is unknown from the data sheet provided by manufacturer. The depth taken into account is shifted depth which is the distance from the object to the lens of the left camera. The unknown distance from the camera lens to sensor plane thus can be compensated. In Fig. 9, the camera measured depth versus shifted depth is illustrated.

TABLE I

RESULTS OF SINGLE CAMERA AND STEREO CAMERA CALIBRATION

	Error (pixel)
Left camera single calibration	0.456
Right camera single calibration	0.458
Stereo camera calibration	1.106

TABLE II

MEAN ERROR AND STANDARD DEVIATION OF DELTA DEPTH

	Mean, mm	Standard Deviation
Displacement of camera measured depth ΔZ_{m}	4.867	0.63
Error	0.544	0.33

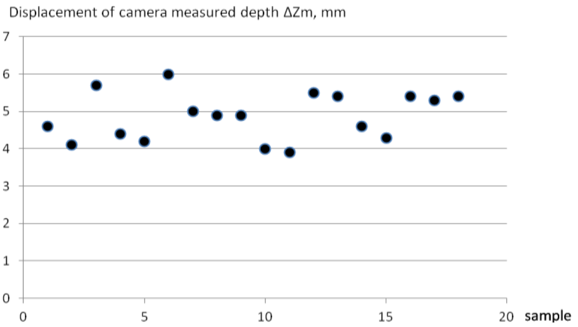

Fig. 10. ΔZ_{m} camera measured depth of all measured samples.

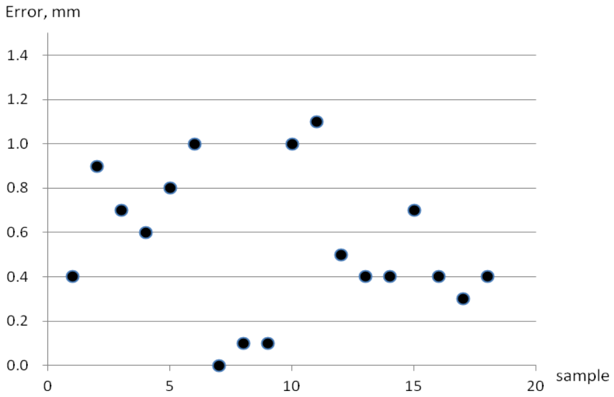

Fig. 11. Camera measured depth versus shifted depth.

The error is defined in such way (14)

$$Error = \left|\Delta Z_{\mathrm{shift}} - \Delta Z_{\mathrm{m}}\right|, \tag{14}$$

where $\Delta Z_{\mathrm{shift}}$ is the displacement of shifted depth between 2 measures, and ΔZ_{m} is the displacement of camera measured depth between 2 measures. Using this error could eliminate the necessity of compensating the distance between the lens and sensor plane. For all the measures in Fig. 9, $\Delta Z_{\mathrm{shift}}$ is 5 mm. In Fig. 10 it indicates the $\Delta Z_{\mathrm{shift}}$. In Fig. 11, the error of all samples is displayed.

As indicated in Fig. 10 and Fig. 11, the error of measured depth could reach around 1 mm. Almost all of the measures fall in the range of around 1 mm. This implies that almost each time when the object is moved 5 mm away from camera, the camera measured depth displacement $\Delta Z_{\mathrm{shift}}$ is within 5 mm ± 1 mm. Table II indicates the mean value and standard deviation of ΔZ_{m} and error.

The depth accuracy depends on several factors. From a hardware point of view, pixel size, camera sensor quality and lens affect the depth calculation. But once the hardware is selected, there is not much one can do with hardware to affect depth accuracy. Other factors that affect depth accuracy are taking place during setting up of the stereo system. Calibration of cameras certainly is important. Normally it is acceptable if reprojection error is around 1 pixel. Another factor is the baseline and depth range. For a certain baseline, the further the distance is from the camera, the less accurately in depth the stereo rig can detect. With the distance increased, the error increases exponentially. If an object is fixed at a certain distance from the cameras, larger baseline leads to a better resolution.

Additionally, different baselines and depth ranges will also lead to different maximum disparity value. The further the object is away from the cameras, the smaller is the disparity value. For a certain distance, a smaller baseline means smaller maximum disparity value. The results of the presented stereo vision are based on a certain baseline and working depth range. If the desired working depth range is different from the one in experiment, the accuracy should be evaluated separately.

V. Discussion

The proposed descriptor and correcting algorithm is proven to work effectively in different lighting conditions. An object is detected from different orientation and distance. Both monochromatic color pattern and randomly selected RGB color pattern which represents un-textured and textured surface are tested. Although in this paper the un-textured surface is not applicable for block matching, it still demonstrates that it works on un-textured surface. For the third object in experiment, it can also work on a non-flat surface.

There are several factors that can affect the accuracy of stereo vision system. Pixel size, calibration of cameras, correspondence matching algorithm will all have effects on depth accuracy. In addition, the depth resolution is different for different depth range. With a proper calibration, the stereo vision in this paper can achieve 1 mm depth error with the pixel size being 3.75 µm, focal length being 6 mm and baseline being 200 mm. The underlying idea of the pose estimation is aligning 2 planes. The plane is extracted from 3 feature points of object. The plane then represents the pose of the object. The absolute pose is not important since the algorithm calculates the relative transformation between the desired pose and current object pose. This pose deviation can be used to control the manipulator.

The results of the paper will be used in the controller part of a visual servoing system. The work presented in the paper will also be further developed and optimized towards being more intelligent. The current training is still supervised by human. It could be developed into an unsupervised training or used as an agent-based solution for teaching. The proposed object recognition algorithm could be investigated more systematically and quantitatively. Moreover, in stereo vision part, several possibilities of improvement could be investigated. The current correspondence matching cannot deal with un-textured surface. Other matching algorithms could be tested. The pose estimation approach is using 3 feature points. To further increase robustness, more feature points could be selected or the use of fusion of different descriptors could be investigated. Occlusion is another problem. It will cause problems for the visual servoing part. Further techniques and algorithms may be applied to solve it.

References

[1] F. Chaumette and S. Hutchinson, "Visual servo control. I. Basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, Dec. 2006. https://doi.org/10.1109/mra.2006.250573

[2] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996. https://doi.org/10.1109/70.538972

[3] E. Malis, F. Chaumette, and S. Boudet, "2 1/2 D visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 2, pp. 238–250, Apr. 1999. https://doi.org/10.1109/70.760345

[4] C. Peng and J Krumm, "Object recognition with color cooccurrence histograms," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999*. vol. 2, 1999. https://doi.org/10.1109/cvpr.1999.784727

[5] C. Ancuti and P. Bekaert, "SIFT-CCH: Increasing the SIFT distinctness by Color Co-occurrence Histograms," *2007 5th International Symposium on Image and Signal Processing and Analysis*, Sep. 2007. https://doi.org/10.1109/ispa.2007.4383677

[6] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010. https://doi.org/10.1109/tpami.2009.154

[7] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. https://doi.org/10.1023/b:visi.0000029664.99615.94

[8] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008. https://doi.org/10.1016/j.cviu.2007.09.014

[9] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," *2011 International Conference on Computer Vision*, Nov. 2011. https://doi.org/10.1109/iccv.2011.6126542

[10] O. Wasenmüller and D. Stricker, "Comparison of Kinect V1 and V2 Depth Images in Terms of Accuracy and Precision," *Lecture Notes in Computer Science*, pp. 34–45, 2017. https://doi.org/10.1007/978-3-319-54427-4_3

[11] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 32, no. 5, pp. 922–923, Sep. 1976. https://doi.org/10.1107/s0567739476001873

[12] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 34, no. 5, pp. 827–828, Sep. 1978. https://doi.org/10.1107/s0567739478001680

[13] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb. 1992. https://doi.org/10.1109/34.121791

[14] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-Squares Fitting of Two 3-D Point Sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 5, pp. 698–700, Sep. 1987. https://doi.org/10.1109/tpami.1987.4767965

[15] J. Heikkila, "Geometric camera calibration using circular control points," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1066–1077, 2000. https://doi.org/10.1109/34.879788

Lei Shi obtained his B. Eng in electrical engineering and automation from Shanghai Maritime University, China, in 2006, and M. Sc in mechatronics from Tallinn University of Technology, Estonia, in 2017. Author's research interests include computer vision, sensor fusion and vision based control of robotics.

Address: Sõstra 6-216, Tallinn, Estonia

E-mail: lei_shi_07@yahoo.com