Pose Estimation of Randomly Organized Stator Housings

Thomas B. Moeslund and Jakob Kirkegaard

Laboratory of Computer Vision and Media Technology
Aalborg University, Denmark
E-mail: tbm@cvmt.dk

Abstract. Machine vision is today a well-established technology in industry where especially conveyer belt applications are successful. A related application area is the situation where a number of objects are located in a bin and each has to be picked from the bin. This problem is known as the automatic bin-picking problem and has a huge market potential due to the countless situations where bin-picking is done manually. In this paper we address a general bin-picking problem present at a large pump manufacturer, Grundfos, where many objects with circular openings are handled each day. We pose estimate the objects by finding the 3D opening based on the elliptic projections into two cameras. The ellipses from the two cameras are handled in a unifying manner using graph theory together with an approach that links a pose and an ellipse via the equation for a general cone. Tests show that the presented algorithm can estimate the poses for a large variety of orientations and handle both noise and occlusions.

1 Introduction

Machine vision is today a well-established technology in industry and is becoming more and more widespread each year. The primary area of success for machine vision is conveyer belt applications, e.g., quality control and robot guiding. The latter is the task of providing robots with positioning data for objects located on a moving conveyer belt. Normally a machine vision system is combined with some kind of mechanical device that ensures that only one object is presented to the system at a time, i.e., no occlusion is present.

A related application area is the situation where a number of objects are located in a bin, see figure 1, and each has to be picked from the bin and placed on a conveyer belt in a predefined pose. This problem is known as the automatic bin-picking problem [13]. A general solution to this problem has a huge market potential due to the countless situations where bin-picking is done manually.

Many different approaches to the bin-picking (and related) problems have been suggested. They can be divided into two categories: model-based approaches and appearance-based approaches.

In the *appearance-based* approaches, a large number of images are obtained of the object and these different appearances are then used when pose estimating the objects. The immediate advantage of this is, that the scene data and

the model data are expressed in the same terms together with its capability of handling objects with no apparent features like lines or corners. The disadvantage of *appearance-based* methods is that the appearance of an object is highly dependent on illumination, viewpoint and object pose [14]. For example, in [1] between 4.000 and 12.000 images of different viewpoints are applied to learn the appearance of the pose of a particular object. In [10] a model of the object is represented as a probability distribution describing the range of possible variations in the object's appearance.

The *model-based* approach on the other hand, represents objects through features, their type and spatial relations. The advantage of *model-based* representations is, that they generate compact object descriptions and offer some robustness against occlusion and some invariance with respect to illumination and pose variations. The disadvantage is that the feature representation cannot be compared directly with the intensity images and that a scene feature extraction therefore is needed. For example, in [7] a wire frame of the model is used and compared with edges in the image. In [9] the CAD (Computer Aided Design) model of the object is used together with relevant object feature points pointed out by the user. In [8] distinct corner features are found in two images and triangulated in order to find the pose of the object.

1.1 Content of this Work

Grundfos [15] is one of the world's leading pump manufacturers and produces more than 25.000 pumps per day using more than 100 robots and 50 machine vision systems. Some of the robots and machine vision systems are used in fully automatic bin-picking applications with well organized objects. However, a large number of unsolved bin-picking problem remain and therefore Grundfos are interested in general solutions for handling the bin-picking problem. This paper presents research in this context.

¿From a machine vision point of view the bin-picking problem is extremely difficult due to the very high number of objects potentially occluding each other and changing the illumination locally due to shadows and reflections. To make the problem tractable we reformulate it to be a matter of picking one and only one object from the bin and *then* finding the pose of this isolated object. The idea being that the combined complexity of the two new problems is less than the complexity of the original problem. The latter problem can be handled by showing the picked object to a camera during the flight from the bin to the conveyer belt, and this pose estimation problem of a single known object in a controlled environment can "easily" be solved. What remains is to find a way of picking one and only one object from the bin.

Many of the objects being produced at Grundfos can roughly be considered as having a cube-like shape, i.e., six sides. Our approach is to have different picking tools for the robot corresponding to the different sides of the cube. We then view the problem of finding and picking an object, as a matter of finding one of the sides and then apply the appropriate picking tool. By looking at the object in figure 1 it can be seen, that it has six "sides" where three are similar

(the smooth sides). To find and pick an object having this type of side facing the camera can, e.g., be done using structured light and stereo-vision followed by a vacuum gripping device [2].

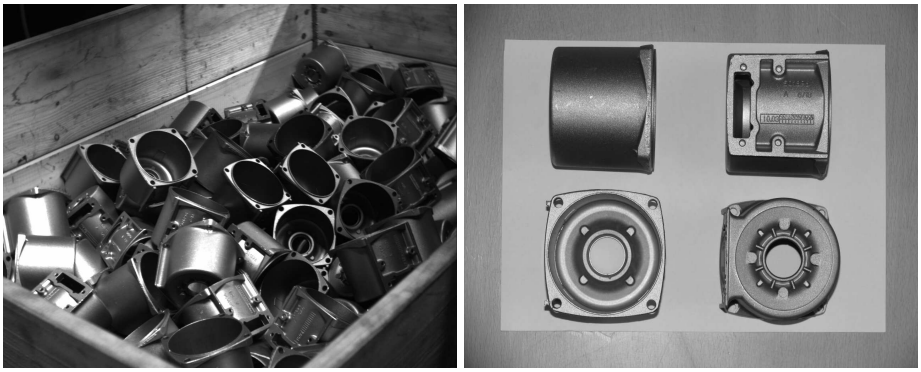

Fig. 1. Left: A bin containing randomly organized Stator Housings. **Right**: The Stator Housing object shown from four different viewpoints on a piece of A4 paper for reference.

In this work we seek a solution to the machine vision problem of finding the "side" representing the opening of such objects, see figure 1, and as a case study we use a bin containing Stator Housings, see figure 1.

The "opening-side" of an object can be characterized by a circle in 3D which projects to an ellipse in an image. Therefore, the problem we are addressing is that of estimating the pose of a circle given it's elliptic projection into two cameras. The paper is structured as follows. In section 2 edge pixels belonging to the same ellipse are grouped and fitted to an ellipse. In section 3 the ellipses found in the cameras are used to estimate the pose of the circles. In section 4 the results are presented and in section 5 a conclusion is given.

2 Estimating the Ellipses

As described above our strategy is to estimate the pose of the objects based on elliptic features in the stereo images. In this section we first describe how the edges extracted from the intensity images are grouped into meaningful segments (denoted *edgels*) each corresponding to an elliptic arch. Secondly, we describe how each edgel is fitted to an ellipse.

Initially we apply the *Canny* edge detector [14] as it not only finds edges, but also groups them into one pixel wide connected segments - edgels, see figure 2. The edgels are then post-processed in two steps in order to ensure that they each contain pixels from one and only one ellipse. The first step is carried out in order to ensure that each edgel only contains two end-points. This is done

by removing crossing point in the edge images, i.e., pixels with more than two neighbors in an 8-connectivity sense, see figure 2.

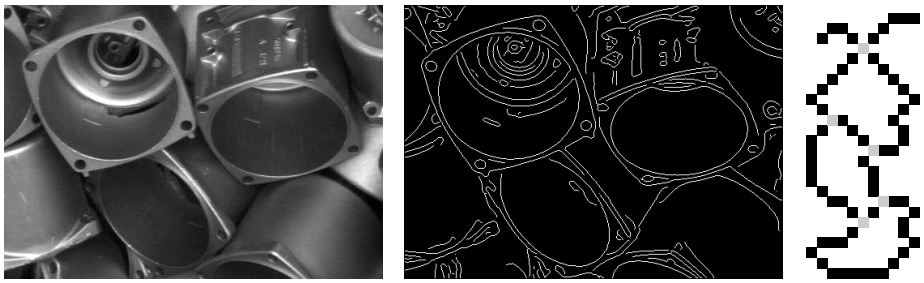

Fig. 2. Left: Small part of an input image. **Middle**: Edge image. **Right**: Crossing points in gray.

As seen in figure 2 edge pixels from different ellipses are often part of the same edgel. The second step therefore removes large concavities by dividing the edge pixels into separate edgels by removing the pixels with a too high curvature. Using standard measures for the curvature turned out to be too sensitive when evaluating the derivatives at particular points. As a result we follow a different approach. Instead of measuring the curvature at one point we filter the curve by dividing it into a number of straight line-segments and then measure the angle between adjacent line segments. We use a modified version of [11].

A typical curve often appearing in the Stator Housing edge images is the one shown in the left most part of figure 3 and clearly has a large concavity at the point of intersection.

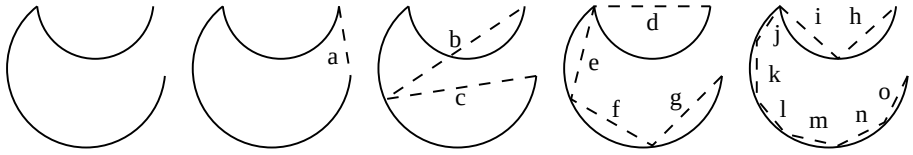

Fig. 3. The principle behind dividing an edgel into straight line-segments.

An edgel is segmented into straight lines by first making the crudest approximation possible, i.e., a straight line connecting the end points. The algorithm then proceeds recursively by dividing each approximating line segment into two line segments at the point of largest deviation, i.e., at the curve point with the largest distance to the approximating line segment.

Each line segment is assigned a significance value, which is the ratio between the length of the line and the largest deviation from the points it approximates. This can be interpreted as, the shorter a line segment is the less deviation is tolerated.

The algorithm continues until the most significant lines are found, i.e., a line segment is only subdivided if it's significance is less than the sum of the significances of the children. Furthermore, if the length of the line or the deviation becomes too small the sub-division is also terminated.

The problem can be posed as a graph search where each node in the graph represents a line segment and has a weight equal to the significance of the line. The tree is then traversed breadth-first and bottom-up searching for nodes that have greater significance than all their children together.

2.1 Ellipse Fitting

After having removed the edge pixels resulting in multiple end-points or large concavities, we are left with a number of edgels[1]. To each of these is fitted an ellipse using the direct fitting method by Fitzgibbon *et al.* [4], which is based on minimizing the algebraic distance between the edge pixels and an ellipse. The method is based on solving a generalized eigenvalue problem constructed from the edgel's points for obtaining the optimal ellipse parameters, i.e., a closed-form solution is obtained.

This algorithm results in a number of ellipses some of which might be very similar. The reason being that one ellipse might be represented as a number of edgels due to noise. We therefore find the ellipses which are similar and merge the corresponding edgels into a new set of pixels which is used to find the parameters for the joint ellipse. The similarity measure is based on a box classifier in the five dimensional space spanned by the ellipse parameters. Finally ellipses with unrealistic parameters are ignored.

3 Pose Estimation

Given a number of ellipses estimated in each image we are now faced with the problem of calculating the corresponding 3D circles. We apply the approach described in [12] where the idea is to find a general cone that both embodies the ellipse and the corresponding 3D circle with radius of the Stator Housing openings. This approach works well but it has the same problem as similar algorithms, namely that two different 3D circles correspond to the same ellipse in the image (and the same general cone). See [5] for details. We therefore need to validate which of the two is the correct solution.

3.1 Circle Pose Validation

In order to solve the validation problem we apply an ellipse matching procedure between the camera frames in order to select the correct circle candidate.

[1] Note that edgels with less than 30 [*pixel*] are ignored altogether.

The problem is posed as a graph search problem as shown by the *association graph* in figure 4. A *node* in the graph represents a *match* between an ellipse in the left and right frame (e.g., 1,3 indicates a match between the 1st left and the 3rd right ellipse), while an *edge* between two nodes in the graph indicates compatibility between two matches.

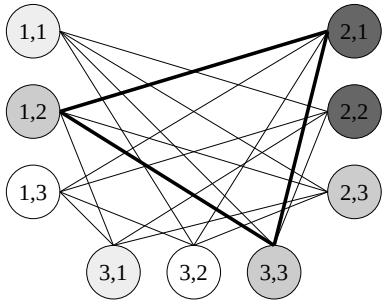

Fig. 4. A graph representing the problem of matching a set of ellipses between the camera frames. Darker node fillings indicate higher weighted nodes and the bold lines indicate the maximally weighted clique.

The association graph is build by creating a node for every pair of ellipses from each camera frame. Many of these nodes can immediately be discarded by investigating the ellipse parameters, while the remaining nodes are given a weight according to $1/(c_{dist} + a_{dist} + b_{dist} + i_{dist})$, where the denominator accumulates distances between the horizontal ellipse centers (c_{dist}) together with differences in major axes (a_{dist}), minor axes (b_{dist}) and inclination angle (i_{dist}) of the ellipse parameters[2].

The problem of finding the best match between the ellipses in the left and right camera frames is then reduced to the problem of finding the maximally weighted set of mutually compatible nodes, i.e., the maximally weighted clique in the association graph [3]. This NP complete problem is approached by stochastic optimization (simulated annealing with linear cooling scheme, [6]) of the function given by equation 1.

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i - \lambda \sum_{i=1}^{n} w_i c_i \tag{1}$$

where $f(\cdot)$ calculates a gain for the given set of graph nodes defined by the membership vector $\mathbf{x}$. The length of the membership vector is given by n and the term w_i states the weight while the binary variable x_i denotes clique membership of the $i'th$ node. The binary variable c_i states whether the $i'th$ node can be part of the clique stated by the membership vector $\mathbf{x}$. The state of c_i for a given node

[2] Note that rectified images are used.

is determined from the inverse graph, i.e., for a set of nodes to form a clique, no two clique nodes must be connected in the inverse graph. This is formally stated by equation 2, where E is the edge set for the association graph. The factor λ (set to unity in the current implementation) is included to control the balance between the gain and the penalty.

$$\forall\, i, j \in \overline{E} : x_i + x_j \leq 1 \tag{2}$$

The result of the optimization is a membership vector indicating the nodes in the (approximated) maximally weighted clique (i.e., a number of compatible ellipse matches). Each match states that two pose candidates in the left frame have been matched with two pose candidates from the right frame. The final pose candidate for each match is then chosen by transforming the two right camera frame pose candidates into the left camera frame (stereo rig is assumed calibrated) and then finding which of the two possible combinations that are most similar. The similarity is measured using the distance between the circle centers and the angle between the normal vectors of the intersecting planes.

3.2 Circle Matching and Quality Measure

Having estimated the 3D pose of the different circles is the same as estimating the 5 DoF for the Stator Housings. Before communicating these results to the robot we also have to calculate a quality measure for each circle, i.e., which object should the robot pick first. A high quality object is an object which is not occluded, which is rotated so that the opening is maximum in the direction of the robot, and which is one of the top objects in the bin. The occlusion is measured using Chamfer matching, i.e., we synthesis the estimated pose of the object into the image and count the distance from each projected point to the nearest edge pixel. To avoid the influence of the actual distance of the projected object, the measure is normalized. The second measure is simply expressed as the *cosine* of the angle between the normal vector for the circle and the camera (both in the left camera). The third measure is the ratio between the distance from the circle to the camera and the distance from the camera to the circle closest to the camera. In mathematical terms the quality measure for the $i'th$ circle is

$$q(i) = w_1 \frac{M_i}{\max\left\{M_i, \sum_j \varepsilon_j\right\}} + w_2 \cos(\varphi_i) + w_3 \frac{\delta_i}{\max\{\delta_j\}} \tag{3}$$

where w_1, w_2, and w_3 are weight factors, M is the number of projected points, ε_j is the distance from the $j'th$ projected point to the nearest edge pixel in the image, φ_i is the angle between the normal vector of the circle and the camera, and δ_j is the distance from the center of the $j'th$ circle to the camera. Note that each term in the quality measure is normalized to the interval $[0; 1]$.

4 Results

The first test is done on synthetic data in order to clarify how the angle between the simulated circle normal vector and the view point vector (denoted ϕ in the following) affects the estimated pose. The test is based on 10.000 random circles with realistic parameters. Each circle is projected into the two images and the corresponding 3D circle is estimated. For each reconstructed circle we measure 1) the absolute error between the simulated circle and the estimated circle center and 2) the angle between the simulated circle normal vector and the estimated circle normal vector. Both measures are calculated with or without noise (each pixel is translated with a random value in the range $[-3; 3]$ in both x and y directions), see figure 5.

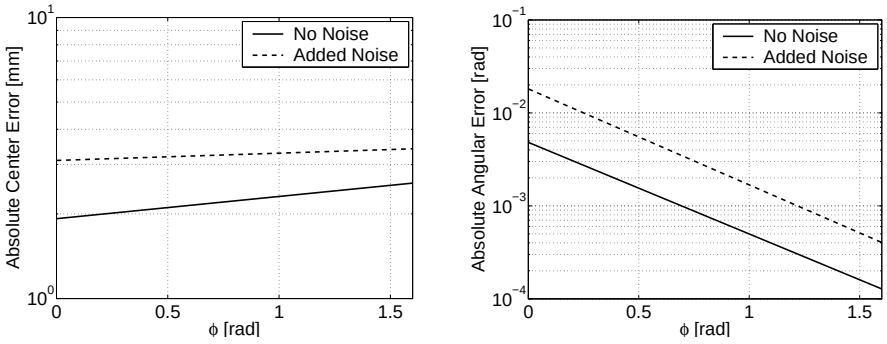

Fig. 5. Errors in the system based on 10.000 randomly generated circles. See text for further details. Note that each data set is fitted to an exponential curve, i.e., a line in a semilogarithmic coordinate system.

When no noise is added the errors between the simulated and estimated centers are around 2 $[mm]$. In the case of added noise, however, the error increases with view point angle. This phenomenon can be explained by the type of noise introduced. As the circle is seen from an oblique angle (view point angle approaching $\pi/2$) the projected ellipse will contain fewer pixels. As the noise are introduced on a pixel level, the particular noise will be most effective when the viewing angle increases.

The circle orientation tests show a tendency towards smaller errors as the view angle increased. This somehow non-intuitive property has been further investigated and the result has been supported by an analytic sensitivity analysis [5].

The second test is a quantitative test where 30 images like the one in figure 6 are used for each camera. We manually judged which circles the system should be able to find (based on visibility in both images) and used this as ground truth. In figure 6 (right) the height of each column is the ground truth. Furthermore

the figure also illustrates the number of false positive and the number of false negatives. The false positives are mainly due to the fact that the opposite side of a Stator Housing contains a similar circle. The false negatives are mainly a result of corrupted edgels due to the illumination or incorrect splitting and merging of edgels.

Recall that we are only interesting in picking one object per time instance, i.e., after the object is removed a new image of the bin might provide a new set of ellipses. Therefore our success criterion is not a good recognition rate in general, but a good recognition rate among the Stator Housings with the best quality measures. For the 30 test examples the objects with the two or three highest quality measures are virtually always pose estimated correct.

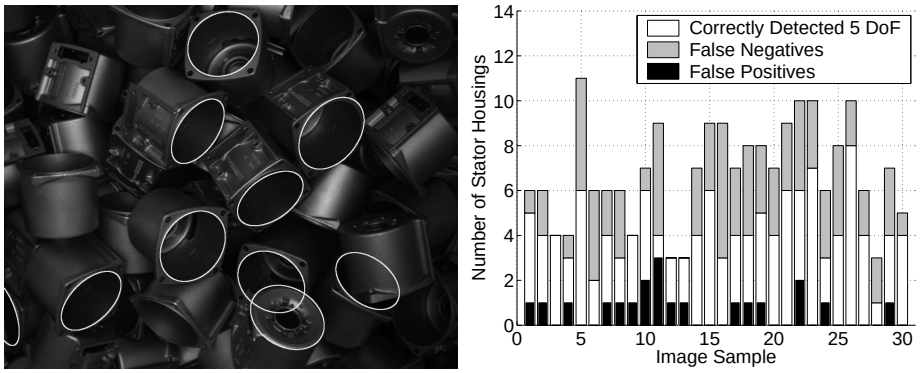

Fig. 6. Left: Estimated 3D circles projected into the left camera image. **Right**: Quantitative test results for the 30 test examples.

5 Conclusion

A general solution to the bin-picking problem has a huge potential in industry. The problem is, however, very difficult and therefore we have reformulated it as matter of picking one and only one object from the bin and *then* finding the pose of this isolated object. The latter task is doable using standard techniques. The former is addressed by the notion of different algorithms and picking tools for each "side" of the object. In this paper we have presented a general solution to pose estimating objects containing circular openings which is a typical characteristic for objects manufactured at Grundfos. The presented algorithm can estimate the poses for a large variety of orientations and handle noise and occlusions primary due to a pose-processing step where information from multiple edge segments in two different camera images are combined. Tests show the approach to be a solid solution to this problem.

Future work includes merging this algorithm with algorithms developed for the other sides of the object and then combining the quality measure for the different methods into one unifying scheme allowing the robot to pick the "best" object at any particular time instance.

References

1. I. Balslev and R.D. Eriksen. From belt picking to bin picking. *Proceedings of SPIE - The International Society for Optical Engineering*, 4902:616–623, 2002.
2. M. Berger, G. Bachler, and S. Scherer. Vision Guided Bin Picking and Mounting in a Flexible Assembly Cell. In *13th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE2000*, New Orleans, Louisiana, USA, June 2000.
3. O. Faugeras. *Three-Dimensional Computer Vision - A Geometric Viewpoint*. The MIT Press, first edition, 1993.
4. A.W. Fitzgibbon, M. Pilu, and R.B. Fisher. Direct least-squares fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):476–480, 1999.
5. J. Kirkegaard. Pose estimation of randomly organised stator housings with circular features. Technical report, Aalborg University, Laboratory of Computer Vision and Media Technology, 2005.
6. S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
7. D.G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.
8. T.B. Moeslund, M. Aagaard, and D. Lerche. 3D Pose Estimation of Cactus Leaves using an Active Shape Model. In *IEEE Workshop on Applications of Computer Vision (WACV)*, Breckenridge, Colorado, Jan 2005.
9. Y. Motai and A. Kosaka. Concatenate feature extraction for robust 3d elliptic object localization. *Applied Computing 2004 - Proceedings of the 2004 ACM Symposium on Applied Computing*, 1:21–28, 2004.
10. A.R. Pope. Learning to recognize objects in images: Acquiring and using probabilistic models of appearance. Technical report, The University of British Columbia, Department of Computer Science, 1995.
11. P.L. Rosin. Nonparametric segmentation of curves into various representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1140–1153, 1995.
12. R. Safaee-Rad, I. Tchoukanov, K. Carless Smith, and B. Benhabib. Three-dimensional location estimation of circular features for machine vision. *IEEE Transactions on Robotics and Automation*, 8(5):624–640, 1992.
13. C. Torras. *Computer Vision - Theory and Industrial Applications*. Springer-Verlag, first edition, 1992.
14. E. Trucco and A. Verri. *Introductory Techniques for 3D Computer Vision*. Prentice Hall, first edition, 1998.
15. *www.grundfos.com*.