

# CATADIOPTRIC MULTIVIEW POSE ESTIMATION FOR ROBOTIC PICK AND PLACE

Markus Heber, Matthias R  ther and Horst Bischof

*Institute for Computer Graphics and Vision, Graz University of Technology, Inffeldgasse 16/II, Graz, Austria*

**Keywords:** Mirror symmetry, Object pose, Contour matching, Industrial application.

**Abstract:** Robotic handling of objects requires exact knowledge of the object pose. In this work, we propose a novel vision system, allowing robust and accurate pose estimation of objects, which are grasped and held in unknown pose by an industrial manipulator. For superior robustness, we solely rely on object contour as a visual cue. We address the apparent problems of object symmetry and ambiguous perspective by acquiring multiple views of the object cheaply and accurately, through a mirror system. Self-calibration of the mirror setup allows us to model the mirror geometry and perform metric multiview contour matching with a known 3D model.

## 1 INTRODUCTION

Automated robotic handling processes rely on exact knowledge of type and pose of the object to manipulate. So the problem of pose estimation is concerned with determining object position and orientation relative to a reference coordinate frame. We especially address the case of uniformly textured, opaque or specularly reflecting objects. Approaches based on 3D reconstruction will fail due to robustness problems. Here, the only robust geometric cue is the object contour, which can be segmented even on transparent objects with specialized illumination. If the 3D model is known a priori, contour shape matching and registration techniques are the method of choice, to avoid laborious appearance teaching.

An early pose estimation approach was introduced by Phong et al. (Phong et al., 1995). It is based on line and point correspondences between images of an object of different pose. The six extrinsic parameters, represented by dual quaternions, are estimated by minimizing a quadric error function. Their minimization technique is compared with the Newton method, and Levenberg-Marquardt optimization. Pose estimation results are compared to ground truth data, as well as the results of Faugeras and Toscani (Faugeras and Toscani, 1986).

Byne and Anderson (Byne and Anderson, 1998) introduced a CAD-based method. They combine geometric descriptions of a 3D model, appearance information and functional information. Online, they gen-

erate hypotheses from these models, based on either edge information, or classification of surface material type. Final pose refinement is done by maximizing a fitting score.

An extensive review of pose estimation methods is given by Rosenhahn et al. (Rosenhahn et al., 2004). In (Rosenhahn and Sommer, 2004) Rosenhahn and Sommer introduced a free-form surface based approach, where surface models are represented by three Fourier descriptors. They estimate the corresponding 3D silhouettes and refine the pose using the iterative closest point algorithm (ICP), introduced by Zhang (Zhang, 1994). In a subsequent work, Rosenhahn et al. (Rosenhahn et al., 2006) compared ICP and a variational method for shape registration via level sets. Evaluation results suggest, that the variational method is more robust against large pose variations, while ICP is more accurate.

A recent approach to CAD-based pose estimation is introduced by Ulrich et al. (Ulrich et al., 2009), where hierarchical views of a CAD model are generated at multiple scale levels. For shape matching they evaluate a similarity measure based on gradient orientation differences. Their experiments show considerable robustness to occlusions, clutter and contrast changes. Chang et al. (Chang et al., 2009) investigated pose estimation and segmentation of specular objects. Specular reflections and specular flow of a known 3D model are used for localization and pose estimation. Experiments show the feasibility of their method under sparse highlights and small environmental mo-

tion.

Most approaches assume the presence of edges or any kind of local features. Furthermore, object symmetries and shape ambiguities are not considered. Our method in contrast also works with untextured, transparent and shiny objects. They can also have smooth surface geometry which is not easily approximated by polyhedral models. In our work, we rely only on object contour information. To overcome the problem of contour symmetry and ambiguous views, we propose a multi-mirror system in combination with a single camera and light source. The result is a cheap, perfectly synchronized multiview system, which can be self-calibrated from any known reference object. Furthermore, our pose estimation procedure is insensitive to local minima, because an exhaustive search over the space of object contours is performed.

## 2 CATADIOPTRIC GEOMETRY

A central perspective camera projection matrix is given by a  $3 \times 4$  matrix  $\mathbf{P}$ , computed from camera calibration matrix  $\mathbf{K}$ , which includes the five intrinsic camera parameters, rotation  $\mathbf{R}$  and translation  $\mathbf{t}$ :

$$\mathbf{P} = \mathbf{K}[\mathbf{R} | \mathbf{t}] \quad (1)$$

A 3D plane is defined as:

$$\mathbf{n}^T \mathbf{x} + d = 0, \quad (2)$$

with normal vector  $\mathbf{n}$  and the distance to the origin  $d$ . Reflections in 3D space are Euclidean transformations, which additionally perform orientation changes. Algebraically, a reflection is given by a matrix  $\mathbf{D}_{4 \times 4}$ , which is able to reflect points  $\mathbf{x}$ , planes  $\Pi$  and cameras  $\mathbf{P}$  over a reflection plane:

$$\mathbf{x}' = \mathbf{D}^T \mathbf{x}, \quad \Pi' = \mathbf{D}^{-1} \Pi, \quad \mathbf{P}' = \mathbf{P} \mathbf{D}^T. \quad (3)$$

Catadioptric devices use reflective devices in a camera's field of view to capture an object from more than one viewpoint. Additionally, catadioptric stereo has geometric and radiometric advantages. One geometric advantage is the reduced number of camera parameters. One radiometric advantage is the replication of light sources due to mirror reflections. The relation between the real camera and its virtual reflection is defined by the mirror reflection matrix  $\mathbf{D}$ , which is defined by the mirror normal  $\mathbf{n}$ , the camera-mirror-distance  $d$  and the camera coordinate frame origin  $\tilde{\mathbf{c}}$  (Gluckman and Nayar, 2001):

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} - 2\mathbf{n}\mathbf{n}^T & \tilde{\mathbf{c}} - 2d\mathbf{n} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix}. \quad (4)$$

A virtual camera is computed from  $\mathbf{P}^{\text{real}}$  as

$$\mathbf{P}^{\text{virtual}} = \mathbf{P}^{\text{real}} \mathbf{D}^T. \quad (5)$$

## 3 POSE ESTIMATION

Pose estimation is based on matching measured contours from all mirror views against a set of pre-generated synthetic contours. Using a known 3D model, and camera-mirror geometry as obtained from calibration, the object is rendered in different poses. From each rendered image, contours are extracted and added to a database. The set of all rendered images covers the space of possible object orientations in front of the camera, sampled in discrete intervals. Pose estimation subsequently is reduced to an exhaustive search within this database. The classification result is guaranteed to be globally optimal with respect to the discretized space of orientations.

### 3.1 System Calibration

We assume the camera projection center to be located at the world coordinate origin. Hence, camera rotation  $\mathbf{R}$  is the identity matrix and translation  $\mathbf{t}$  is zero. Intrinsic calibration is performed as proposed by Zhang (Zhang, 1999). The mirror calibration procedure used within our approach is based on the work of Hu et al. (Hu et al., 2005), where first the mirror plane normal  $\mathbf{n}$  is estimated, followed by camera-mirror-distance  $d$ . For computation of  $\mathbf{n}$ , two pairs of corresponding points between real view and each mirror view are required. These correspondences are obtained via the object convex hull in the real and mirror views. There are exactly two lines, which are tangent to both convex hulls. They are called *limitation lines* and provide a pair of corresponding points each. Furthermore, their intersection provides the vanishing point  $\mathbf{vp}$  of  $\mathbf{n}$ , which coincidentally describes the epipole  $\mathbf{e}$  of the virtual camera. Mirror normal  $\mathbf{n}$  is computed by evaluating the direction of the viewing ray through  $\mathbf{e}$  in the image:

$$\mathbf{n} = (n_x, n_y, n_z)^T = \mathbf{K}^{-1} \mathbf{e}. \quad (6)$$

Camera-mirror-distance  $d$  is computed with knowledge of a single object point  $(x, y)$  and its mirrored correspondence  $(x', y')$ :

$$d = \frac{\Delta u z_0}{2(u' n_z - n_x)}, \quad (7)$$

where  $(u, v)$  are normalized image coordinates, and  $\Delta u = u' - u$ . The nominal distance between camera center and 3D world point  $z_0$  is set to 1, which results in a system calibration up to an unknown scaling factor (Hu et al., 2005). In the case of multiple mirrors, these correspondences cannot be uniquely determined over all views. Hence, an additional point correspondence is established by evaluating the centroid of a

sphere. With known sphere radius, the calibration can be upgraded to metric, including exact scale.

### 3.2 Contour Representation and Similarity Metric

The matching process itself needs to be fast and accurate, but it does not have to be scale- or rotation invariant. We therefore choose a simple approach, where each contour is represented by a set  $S_i$  of neighboring contour points  $S_i = \{\mathbf{x}_{i1} \dots \mathbf{x}_{in}\}$ . Without scale invariance, identical contours have approximately the same number of points, and a similarity metric is computed, using the sum of squared distances of corresponding points:

$$e_{ij} = \sum_{k=1}^n |\mathbf{x}_{ik} - \mathbf{x}_{jk}|^2, \quad (8)$$

where  $e_{ij}$  is the similarity error between two contours  $S_i$  and  $S_j$ .

### 3.3 Contour Extraction and Matching

Shape matching is an exhaustive search for the best matching synthetic contour in all views. The camera intrinsics, mirror parameters, and a contour database, which stores the synthetic object contours for all orientations, are given. Experiments on synthetic contour matching as well as real object pose estimation (see Section 4) have shown that it is sufficient to consider a discrete set of orientations. Considering the space of all object orientations as the set of all roll, pitch and yaw-angles  $(r, p, y)^T, r \in [0^\circ, 360^\circ], p \in [0^\circ, 360^\circ], y \in [0^\circ, 360^\circ]$ , we discretize in  $12^\circ$  steps and have 27k database entries. Selection of the discretization step depends on the underlying application as well as available hardware. Modern graphics cards allow rendering of six views of a detailed 3D model at over 100fps, and a database of 27k entries is generated in roughly five minutes. To estimate an object pose, the contours are extracted from an image and compared to entries in the database, according to the metric in Section 3.2. To further speed up the matching process, contour features like length and aspect ratio are used to reject dissimilar contours early on.

## 4 EXPERIMENTS

We focus on object symmetries and ambiguities, because these are the most challenging cases. In Figure 1 a typical symmetry case is shown. Our experimental hardware setup consists of a monochrome

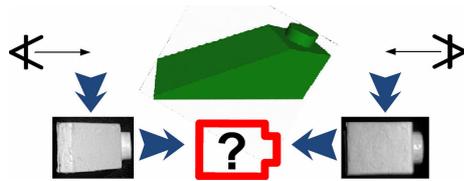


Figure 1: Top views of a sample test object. Extracted image object contours are similar due to object symmetry.

CCD camera, and five planar mirrors. The mirrors are placed transversely in front of the camera. A LED light source illuminates the object coaxially against defined background to simplify the contour extraction process. Due to the transversal placement of the mirrors, light sources are replicated, which results in approximately diffuse illumination conditions and avoids shading on round object borders.

An object is placed in unknown orientation inside the catadioptric setup. We evaluated our setup with three different test objects: (a)  $5 \times 10 \times 5$ mm block, providing a front-back symmetry, if only image contours are taken into account, (b)  $5 \times 15 \times 5$ mm slanting block, providing an upside-down symmetry, and (c)  $10 \times 10 \times 5$ mm slanting block, also providing an upside-down symmetry.

### 4.1 Self Calibration

The system is calibrated as described in Section 3.1. We evaluated the reprojection error (RE) and used it as a measure of calibration accuracy. Over several calibration runs we achieved an average RE of 0.01px. This comes up to a geometric error of  $2\mu\text{m}$  at a camera object distance of 350mm.

### 4.2 Pose Estimation

Pose estimation has been evaluated on synthetic contours as well as real objects, with a focus on symmetries, that cannot be resolved from a single view. These include upside-down flips (*uds*) and front-back flips (*lbs*). Additionally, different rotations (*rot*) have been evaluated. The upside-down ambiguity can be resolved with our proposed method. For symmetric blocks like object (a), a front-back symmetry remains, and quadric blocks would lead to four equivalent poses. In order to evaluate correctness, pose estimation has also been evaluated on synthetic data. Synthetic contours with slightly different poses to the contour database poses were generated. According to a discretization step of  $12^\circ$ , the residual error should not be greater than  $6^\circ$  for a correct match. As presented in Table 1, our results lie within this expected range. Numerical results on the real test objects are

given in Table 2. For each object, (*uds*), (*fbs*) as well as different rotations (*rot*) have been evaluated. For object (a), '?' means a successful match, but a front-back symmetry could not be resolved. Figure 2 shows exemplary matching results.

Table 1: Results on pose estimation of randomly generated synthetic contours at a discretization step of  $12^\circ$ .

| Object  | Runs | Avg. rpy Errors                               |
|---|------|---|
| (a)  | 10   | $r = 5.2^\circ, p = 3.4^\circ, y = 3.5^\circ$ |
| (b)  | 10   | $r = 6.0^\circ, p = 3.2^\circ, y = 5.4^\circ$ |
| (c)  | 13   | $r = 4.7^\circ, p = 4.6^\circ, y = 4.5^\circ$ |

Table 2: Results on pose estimation of real test objects, where e.g 1/2 denotes that one out of two runs was correct.

| Object  | uds   | fbs     | rot   |
|---|-------|---------|-------|
| (a)    | 5 / 5 | '?' / 3 | 4 / 4 |
| (b)   | 4 / 5 | 2 / 3   | 3 / 5 |
| (c)  | 5 / 4 | 3 / 2   | 5 / 3 |

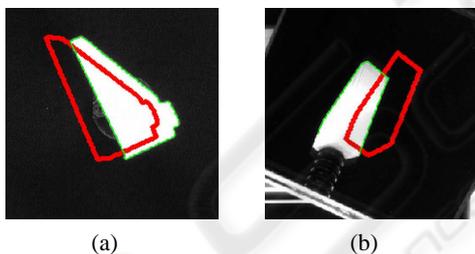


Figure 2: Two examples of contour matching results. Match result are shown with a translational offset for better visualization.

## 5 CONCLUSIONS

We have presented a novel method for object pose estimation. Our approach benefits from employing mirrors in the optical path, due to replicated object illumination, and multiple perfectly synchronized views. We have shown that object symmetry and ambiguous perspective are better resolved than using a monocular setup. The problem of object pose estimation was reduced to an exhaustive search of matching contours. Experimental results show that this procedure does

not get stuck in local minima. Most object symmetries were resolved. Creation and storage of a contour database is feasible up to a certain discretization step. Our choice of  $12^\circ$  might not be sufficient to resolve fine details of some objects, though. Future work includes the intelligent organization of a more dense database by clustering of similar views. Furthermore, evaluation of real objects with given ground truth on their pose will be an issue. Further iterative 3D object registration can also be taken into account. To overcome the discretization error, one could consider further pose refinement, using iterative methods like ICP.

## REFERENCES

- Byne, J. and Anderson, J. (1998). A CAD based computer vision system. *IVC*, 16(8).
- Chang, J. Y., Rsakar, R., and Agrawal, A. (2009). 3D pose estimation and segmentation using specular cues. In *Proc. CVPR 2009, Miami, FL*.
- Faugeras, O. D. and Toscani, G. (1986). The calibration problem for stereo. In *Proc. CVPR, Miami Beach*.
- Gluckman, J. and Nayar, S. K. (2001). Catadioptric stereo using planar mirrors. *Int. J. Comput. Vision*, 44(1).
- Hu, B., Brown, C., and Nelson, R. (2005). Multiple-view 3-D reconstruction using a mirror. Technical report, University of Rochester.
- Phong, T. Q., Horaud, R., Yassine, A., and Tao, P. D. (1995). Object pose from 2-D to 3-D point and line correspondences. *Int. J. Comput. Vision*, 15(3).
- Rosenhahn, B., Brox, T., Cremers, D., and Seidel, H.-P. (2006). A comparison of shape matching methods for contour based pose estimation. *Combinatorial Image Analysis*.
- Rosenhahn, B., Perwass, C., and Sommer, G. (2004). CVOnline: Foundations about 2D-3D pose estimation. *CVOnline*.
- Rosenhahn, B. and Sommer, G. (2004). Pose estimation of free-form objects. In *Proc. ECCV 2004, Part I, T. Pajdla and J. Matas (Eds.)*.
- Ulrich, M., Wiedemann, C., and Steger, C. (2009). CAD-based recognition of 3D objects in monocular images. In *Proc. ICRA 2009, Kobe, Japan*.
- Zhang, Z. (1994). Iterative point matching for registration of free-form curves and surfaces. *Int. J. Comput. Vision*, 13(2).
- Zhang, Z. (1999). Flexible camera calibration by viewing a plane from unknown orientations. In *Proc. ICCV*.