Catadioptric Silhouette-Based Pose Estimation from Learned Models

Christian Reinbacher, Markus Heber, Matthias Rüther, and Horst Bischof

Institute for Computer Graphics and Vision, Graz University of Technology,
Inffeldgasse 16/II, Graz, Austria
{reinbacher,mheber,ruether,bischof}@icg.tugraz.at
http://www.icg.tugraz.at

Abstract. The automated handling of objects requires the estimation of object position and rotation with respect to an actuator. We propose a system for silhouette-based pose estimation, which can be applied to a variety of objects, including untextured and slightly transparent objects. Pose estimation inevitably relies on previous knowledge of the object's 3D geometry. In contrast to traditional view-based approaches our system creates the required 3D model solely from the object silhouettes and abandons the need to obtain a model beforehand. It is sufficient to rotate the object in front of the catadioptric camera system. Experimental results show that the pose estimation accuracy drops only slightly compared to a highly accurate input model. The whole system utilizes the parallel processing power of graphics cards, to deliver an auto calibration in 20 s and reconstructions and pose estimations in 200 ms.

Keywords: pose estimation, model creation, shape from silhouette, catadioptric multi-view.

1 Introduction

Robotic pick & place deals with the problem of automated handling of objects. Typical tasks include sorting, packaging and automated manipulation. In order to correctly place an object, its position and orientation with respect to the actuator has to be known. In some scenarios the orientation is given by the way objects are produced, e.g. filled bottles moved on a conveyor belt, but typically the pose of the object has to be determined on the fly.

Vision-based pose estimation has become popular in industrial settings. Despite the vast amount of literature on various techniques, we limit ourselves to view-based approaches which were quite popular for a time and were recently revisited for industrial problems [1,2,3,16]. Here the object pose is determined by comparing the query image with precomputed 2D reference views of a known 3D model. Removing the translation by normalizing for the object location leads to three unknown degrees of freedom given by the possible rotations of the object. Hence, it is feasible to create reference views by placing a virtual camera on a sphere with the object in its center and later compare an acquired image to these

A. Heyden and F. Kahl (Eds.): SCIA 2011, LNCS 6688, pp. 403–413, 2011.

views. The accuracy of view-based approaches depends on the sampling density of the pose range, and the quality of the 3D model of the object.

Most state of the art methods are designed to work with man-made objects. These objects can be represented by a polyhedral 3D model, and typically resemble sharp edges, which are used as features for the pose estimation. However, there is a large class of objects which have an organic structure, lack edge features and are untextured or slightly transparent (e.g. organic moulding parts). The only remaining cue for them is the filled outline of the object, its silhouette. Using silhouettes, the pose estimation problem reduces to a 2D shape matching problem, where the best matching shape out of a database of precomputed views defines rotation and translation with respect to a camera.

A prerequisite for view-based methods are 3D models, which have to be created by either modeling them in CAD systems or by scanning the object. Reconstruction methods based on Structured Light [14,4] are able to produce very accurate reconstructions even from untextured objects. Shiny surfaces are difficult to handle for structured light methods. However, the object silhouette can also be used for 3D reconstruction.

Structure-from-Silhouette (SfS) methods use only a number of silhouettes to produce an approximation of the 3D object called the Visual Hull (VH), originally introduced in [9]. A visual hull is guaranteed to contain the object but it can be a coarse approximation depending on the number of cameras observing the object. SfS methods typically require a calibrated camera setup, which is able to capture the object from several defined viewpoints. A very elegant and at the same time inexpensive approach is to use a catadioptric system. Reflective surfaces like e.g. mirrors are placed in the field of view of a camera, to create additional views of a target object in a single image. Each mirror creates an additional virtual camera with viewpoint behind the mirror surface. Catadioptric camera setups have appealing advantages over conventional multi-view camera systems: a) they allow for a cheap and perfectly synchronized multi-view setup with a single camera and b) they reduce the number of camera parameters [5].

We propose to use a catadioptric camera system with planar mirrors for both, model learning and model-based pose estimation. The whole work flow of camera calibration, model creation and refinement, and finally pose estimation solely relies on the silhouettes of the objects. Our contributions are twofold: first, we propose to use a visual hull representation of an object as input to our model-based pose estimation. Second, we build an integrated system which can be used for pose estimation as well as model creation, needed by the pose estimation method.

Our experiments give evidence, that the proposed approach works for a variety of objects, where traditional approaches based on image features clearly fail.

2 Method

Our pose estimation method consists of four parts: a) silhouette extraction, b) system calibration, c) model creation, and d) pose estimation. In the following sections we will present the multi-view system, and how it is calibrated using

silhouette information only. During this calibration a 3D model of the object is implicitly created. This model is further refined and subsequently used as input for our pose estimation method. All methods rely on one and the same hardware setup and do not require additional helper devices.

2.1 Silhouette Extraction

The accurate extraction of the object outline is crucial, since it directly influences the accuracy of all subsequent calibration and reconstruction steps. In this work we decided to adopt a recently proposed variational segmentation method introduced in [13]. Segmentation is performed by minimizing the energy:

$$\inf_u \left\{ \int_\Omega g\,|\nabla u|\,\mathrm{d}x + \lambda \int_\Omega u f \mathrm{d}x \right\}, \tag{1}$$

with $u : \Omega \to \{0, 1\}$. The function $g(x)$ is an edge indicator function which is low at a strong edge and high in homogeneous regions. The user-provided potential function f represents the likelihood of every pixel to belong to foreground or background, respectively.

We define $f = -|I - I_{\text{background}}|$ and $f = \infty$ at the image border, because both, background and foreground regions have to be given as input to the segmentation method. In contrast to background subtraction, we perform segmentation with an additional edge term and a powerful smoothing prior, which filters out segmentation errors caused by noise.

2.2 Catadioptric Camera Setup and Calibration

Our setup consists of a single camera, a light source, and n planar mirrors. Figure 1 shows a cross section of the camera setup. The robotic actuator moves the object between the radially arranged mirrors, such that the object is visible in every mirror.

It was shown by Hu et al. [8] and later by Heber et al. [6] that a catadioptric system with planar mirrors can be calibrated solely from outlines of objects within the mirror setup. The method requires an intrinsically calibrated camera $\mathbf{P_{real}}$ at the origin of the world coordinate frame.

The calibration of the catadioptric system results in a set of projectionmatrices

$$\mathbf{P_i} = \mathbf{P_{real}}\,\mathbf{D_i^T}, \tag{2}$$

where $\mathbf{D}_{4\times4}$ defines a reflection matrix, corresponding to a planar mirror in 3D space:

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} - 2\mathbf{n}\mathbf{n}^T & \widetilde{\mathbf{c}} - 2d\mathbf{n} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \quad . \tag{3}$$

Reflections in 3D space are Euclidean transformations, which additionally perform orientation changes. They depend on mirror plane normal $\mathbf{n}$, camera-mirror-distance d, and camera coordinate frame origin $\widetilde{\mathbf{c}}$ as proposed by Gluckman et al. [5]. For details on recovering plane normal and camera-mirror-distance we refer the reader to [6].

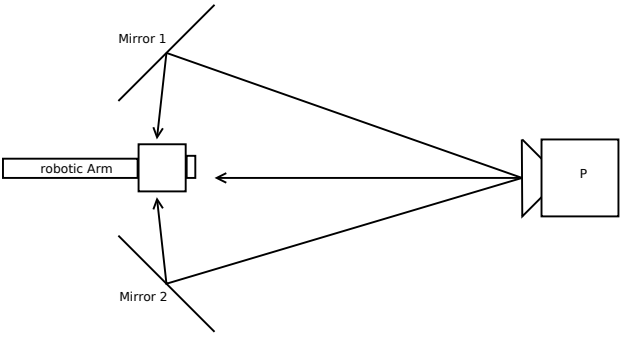
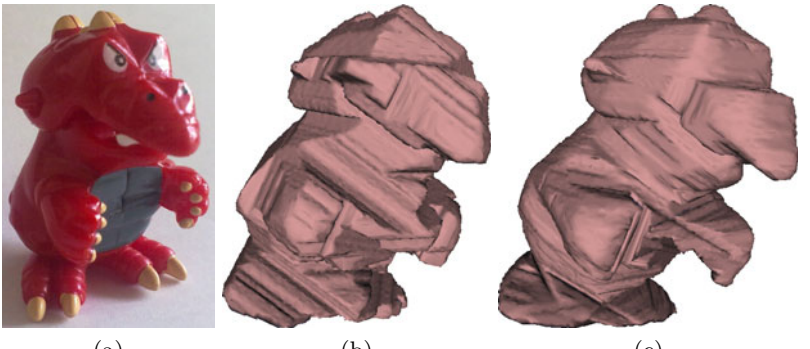


Fig. 1. Schematic of our catadioptric setup. The mirrors are arranged radially around the object.

Having these camera projection matrices along with a set of silhouettes, we are able to estimate a coarse approximation of the object by computing the visual hull. The visual hull is defined as the intersection of all viewing cones, that are generated via back projection of the 2D silhouettes into 3D space.

Figure 2 shows two visual hull 3D models of a toy figure, generated from 6 and 30 camera views, respectively. Obviously, increasing the number of views allows to reconstruct more details. However, accurate reconstructions of deep concavities are not feasible with a visual hull approach which poses no problem as our approach solely relies on object silhouettes.



(a) (b) (c)

Fig. 2. Comparison of visual hull reconstructions of a toy figure from (b) 6 and (c) 30 camera views, respectively

2.3 Reconstruction Refinement

As shown in Section 2.2, the initial model can be refined by adding more views. A cheap way to add new views is given by moving either the object or the camera

setup. In either way the relative orientations of all cameras have to be known. In our application, a robotic end-effector anyway holds the object within the camera setup for further placement, so we rotate the object around the last joint in front of the camera. Each rotation introduces $n+1$ new virtual cameras. The motion of the new set of cameras with respect to the original ones is restricted to a circular motion around a common rotation axis. We consider this motion unknown and implicitly calibrate it during reconstruction.

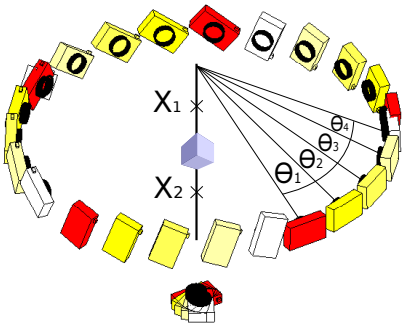


Fig. 3. Positions of the virtual cameras around the object and parametrization of the camera movements. Cameras of the same color belong to one turn of the object in front of the camera setup.

We parameterize the motion with $k+2$ parameters for k movements of the object: the rotation axis, defined by 2 points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^3$, and k angles $\theta_1 \dots \theta_k$ with respect to the original camera position. The coordinate system is defined, such that the real camera is located at the origin and the object at $[0, 0, 1]$. Without loss of generalization, we are able to reduce the dimensions of $\mathbf{x}_1, \mathbf{x}_2$ from $\mathbb{R}^3$ to $\mathbb{R}^2$ by fixing their z coordinate to lie on a plane in front of and behind the object, respectively. Since our fully calibrated $n+1$ camera setup is rigid, we can parameterize $(n+1)(k+1)$ cameras by only $k+2$ parameters.

A setup with $n = 5$ mirrors and $k = 4$ rotations yields a 30 camera multi-view system and can be parametrized by 6 parameters. This example can be seen in Fig. 3.

To automatically determine these parameters, the concept of silhouette consistency for auto-calibration was introduced in [7], which we will briefly discuss here. Given a set of silhouettes $\mathbf{S_i}, i \in [1, k]$ and its corresponding camera parameters $\mathbf{P_i}, i \in [1, k]$, the goal is to maximize the coherence of the measured silhouettes and the model projections. Every optic ray, defined by the camera center and a silhouette pixel in one view, must intersect the silhouette in any other view. This holds for perfect segmentation and camera calibration. Due to noise in both, segmentation and camera position, the above constraint will not hold for some rays.

Hernandez et al. proposed a simple metric to measure the degree of consistency for a set of $(\mathbf{S_i}, \mathbf{P_i})$ by simply counting the number of pixels, that do not comply

to the consistency property. This metric can be computed by creating a visual hull, defined by the silhouettes and current projection matrices, back-projecting it into the camera images and comparing the two silhouettes. For a visual hull V defined by $(\mathbf{S_i}, \mathbf{P_i})$ and its projection into an image $\mathbf{S_i^V}$, we use the ratio of the areas between $\mathbf{S_i}$ and $\mathbf{S_i^V}$ as a consistency measure:

$$C(\mathbf{S_i}, \mathbf{S_i^V}) = \frac{\int \left(\mathbf{S_i} \cap \mathbf{S_i^V}\right)}{\int \mathbf{S_i}} \quad . \tag{4}$$

In order to find optimal camera positions, we seek to maximize the total silhouette consistency

$$\sum_i C(\mathbf{S_i}, \mathbf{S_i^V}) \quad . \tag{5}$$

The problem is solved by the derivative-free Nelder-Mead algorithm [11]. During the optimization process, a visual hull approximation is being built implicitly. The evolution of the model can be seen in Fig. 2 for the initial camera setup and after 4 rotations of the object.

For the visual hull we use a simple volumetric space carving approach, originally proposed by [9]. Since the optimization procedure invokes the visual hull creation many times, we implemented a very efficient simple space carving method [15], that utilizes the parallelism of modern graphics cards. The resulting voxel model is then transformed into a triangulated mesh by applying a standard marching cubes algorithm proposed by [10].

2.4 Model-Based Pose Estimation

With a 3D model at hand, we seek to determine the rotation (roll, pitch, yaw) of an object with respect to the camera system. Seeing that we can only use the object boundaries as input we chose to extend the approach of [12] to a multi-view setup. There the authors did pose estimation by comparing the outline of an object to a database of reference views, created from a 3D model of the object. This potentially large database is indexed into a hierarchical structure by finding similar views, and by grouping them together in a bottom-up fashion. A rotation-invariant match to the database yielded pitch and yaw. The roll angle is determined by the matching procedure.

The original method was designed to be used in a single-view setting. To incorporate the remaining views created by the planar mirrors, we propose to apply the algorithm to only one silhouette. We use the other cameras for verification of the potential matches. Due to the fact, that the roll angle is determined during matching, the additional views can not be stored in the database, but have to be created on-the-fly from the 3D model.

In our setting, the silhouette produced by the real camera is always complete. The views from the virtual cameras may be partially occluded by the robotic end-effector. To cope with this, we propose a simple partial contour matching method. A closed contour from the database is given as a vector of points $\mathbf{C_{db}} = \langle \mathbf{p_1}, \cdots, \mathbf{p_n} \rangle$. The partial query contour is given by $\mathbf{C_q} = \langle \mathbf{q_1}, \cdots, \mathbf{q_k} \rangle$. In order

to match them, we first perform a linear search, where we align $\mathbf{q_1}$ with every point of $\mathbf{C_{db}}$, and measure the Euclidean distance of $\mathbf{q_k}$ to the closest point in $\mathbf{C_{db}}$. This pre-selection yields a set of possible start points for which the final error measure is obtained. The error is given by the sum of squared distances of the aligned $\mathbf{C_q}$ to the closed contour $\mathbf{C_{db}}$.

Out of a list of potential matches, the element of the database with the lowest error defines the rotational part of the object pose. The translation is restricted by the robotic end-effector. Remaining translation errors due to inaccuracies of the robot can be easily determined from the 2D input image.

3 Experiments

In our experiments we use a catadioptric system with 5 planar mirrors, which yields a 6-camera multi-view system. We first evaluate the accuracy of the proposed calibration and reconstruction method. In order to verify our claim that visual hull reconstructions can be used for pose estimation, we use synthetic images of previously scanned 3D models obtained by a laser scanner. Finally we apply our method to various real-world objects.

3.1 Calibration Accuracy

Intrinsic camera calibration is performed using the method proposed by Zhang [17]. The extrinsic parameters of the real and virtual cameras are obtained as described in Section 2.2. We evaluated the re-projection error of the calibration sphere center over 5 calibration runs with an average of 0.01 px, resulting in a geometric error of 2 μm at an object distance of 300 mm.

The optimization procedure converges after roughly 200 iterations. The total time for calibration, reconstruction and triangulation of the voxel representation is roughly 1 minute on a quad core PC with 4GB RAM and a NVidia GeForce GTX285. Due to the high repeatability in positioning of the robotic arm, the calibration has to be done only once. Consecutive reconstructions can be carried out in 100 ms for a voxel size of 512^3.

First, we evaluate the accuracy of the estimated rotation angles. The ground-truth is provided by a turn table with a precision of $1/77°$. The average deviation for several runs was $0.37°$.

Second, we evaluate the accuracy of the estimated rotation axis. To get a ground-truth, we let a sphere rotate off-axis, triangulate the center points and fit a plane through the reconstructed 3D points. The normal of the plane is defined as rotation axis. The average angular deviation for several runs was $0.34°$.

3.2 Pose Estimation with Synthetic Images

In the experiments so far we have focused on the quality of the calibration, which directly influences the quality of the reconstructed objects. Now, we want to

Table 1. Results of the synthetic view experiment. 800 synthetic contours generated from the ground-truth (GT) model are queried against differently detailed reconstructions, obtained from the real object. The table shows the mean deviation from the estimated viewpoint to the ground-truth.

Object	VH_6 model			VH_{30} model			GT model		
	r [°]	p [°]	y [°]	r [°]	p [°]	y [°]	r [°]	p [°]	y [°]
Brick 1	11.28	10.79	7.85	8.52	8.26	5.86	4.57	4.08	2.52
Brick 2	6.74	6.44	4.31	5.03	5.21	3.23	2.44	2.78	1.88
Toy Figure	23.54	25.12	21.94	9.40	7.98	5.18	4.17	3.71	2.31

investigate how the reconstruction quality affects the pose estimation accuracy. In order to give a quantitative evaluation, we use two objects for which an accurate 3D model is given.

Each object is reconstructed with the methods presented in Sections 2.2 and 2.3, resulting in two models. The first model is a visual hull reconstruction with 5+1 camera views, the second model was created by turning the object $k = 4$ times, which equals a reconstruction from 30 camera views. Both models are converted into a triangulated mesh. One object used, a toy figure, is depicted in Fig. 4. (b) and (c) show a comparison between a visual hull reconstruction from 30 views and a scanned 3D model from roughly the same viewpoint.

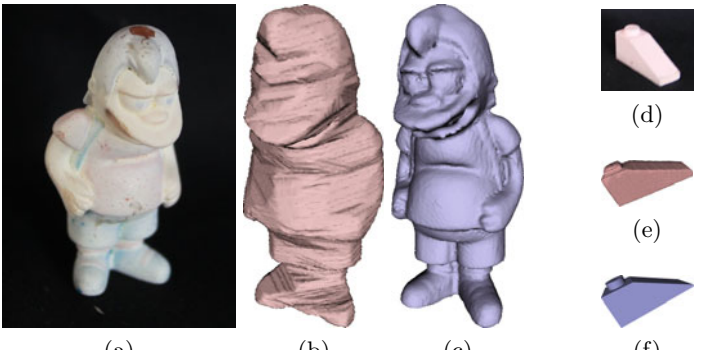


Fig. 4. Comparison of a model generated with our visual hull base reconstruction (b), (e) and a laser scanner (c),(f)

First, we use the ground-truth model to create artificial images of the object. The calibration parameters allow us to simulate the catadioptric camera system, whereas the 3D model gives us ground-truth poses. In our experiment we created 800 views of the object from random viewpoints. A viewpoint is defined by a point on a sphere with the object in its center and a roll angle along the camera's optical axis

$$\mathbf{V_l} = \langle \mathbf{p_l}, r_l \rangle \, , \mathbf{p_l} \in \mathbb{R}^2, r_l \in [0, 2\pi] \quad . \tag{6}$$

For each 3D model of the object a reference view database consisting of 1000 views is created with the method described in Section 2.4. Table 1 represents the results of this experiment in terms of mean deviation of the estimated pose to the ground-truth pose, in roll, pitch and yaw.

Clearly, a more detailed visual hull improves the accuracy of the estimated poses, bringing it very close to the results, that can be obtained by using a laser-scanned model. For the geometrically very complex toy figure, a visual hull obtained from a few camera views is not able to accurately represent the true object, resulting in pose deviations up to 25°. Adding additional views by turning the object in front of the camera improves the result to 8°.

Figure 5 depicts the number of correct pose estimations, given a maximum angular deviation for a simple and a complex object, respectively. The simple model can be approximated quite well even by a coarse visual hull reconstruction, whereas for the complex model the incorporation of more camera views leads to large improvements.

When allowing a maximal deviation from the true pose of 8° our method decreases by 5% for the brick with its simple geometry. For the more complex model, the accuracy decreases by 12%.

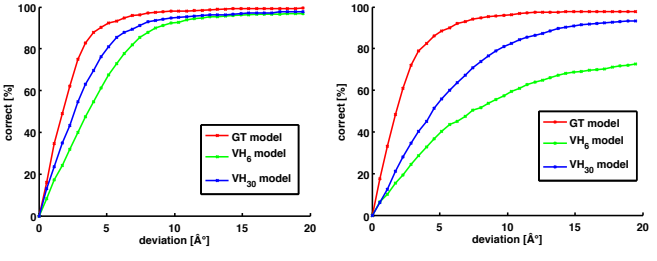


Fig. 5. Results of our synthetic view experiment. The number of correct matches for a given maximum angular deviation is shown for three different 3D models: visual hull reconstructions from 6 and 30 views, respectively, and a laser scanned model. Two different objects were employed: (a) a brick with low complexity, and (b) a toy figure with rather high complexity

3.3 Pose Estimation with Real Images

We validated our approach for a variety of real world objects for which no 3D model was available. For this experiment we used a 2 MP FireWire camera and 5 planar standard mirrors. Figure 6 depicts those objects along with the obtained reconstructions and success rates. Since no ground-truth in terms of correct viewpoints is available, only qualitative results in terms of 'visually correct' or 'visually incorrect' are given. For every object, approximately one out of ten pose estimation was classified as incorrect, nevertheless the failure cases typically were near the correct solutions.

The objects resemble a variety of geometric complexities. They all share the property of being very low textured. The pose estimation results indicate, that our method can be used for almost any objects, as long as their outline can be extracted. A single pose estimation can be obtained in 200 milliseconds, including the time for segmentation, and pose estimation.

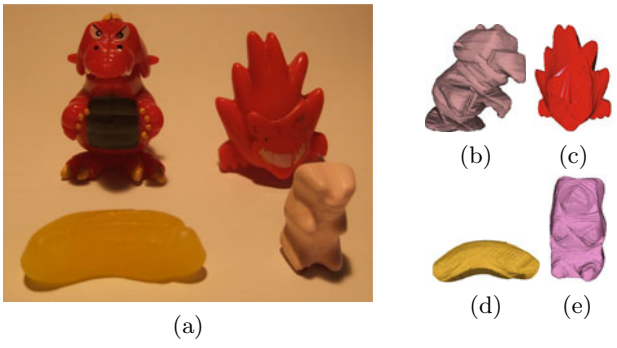


Fig. 6. Objects used for the experiments with real images. (b)-(e) show the obtained reconstructions using the method described in Section 2.3. The success rates of the pose estimation are (b) 83.3%, (c) 90%, (d) 92.8 % (e) 90% respectively.

4 Conclusion

In this work we tackled the problem of pose estimation in the context of robotic pick & place. We introduced an integrated system for model-based pose estimation, without the need of obtaining a model beforehand. Models of new objects are learned on the fly by placing them in front of the camera system. We presented a catadioptric multi-view system, which offers a cheap way of creating several viewpoints with a single camera. The whole process of camera calibration, 3D reconstruction, and pose estimation is solely based on outer contours of the objects. Those silhouettes can be extracted reliably for a variety of objects, making our method applicable to a wide range of products.

We have shown that the visual hull reconstruction can be used for accurate pose estimation, if enough camera views contributed to the reconstruction. Experiments with both synthetic and real objects prove, that the proposed system can be used for objects with arbitrary geometry and surface structure. The implementation of the core algorithm on modern graphics cards allows for pose estimations in 200 ms and system auto-calibration in less than 20 s without user interaction.

Future work may include the investigation of methods tolerant to segmentation errors in order to apply the method to applications with uncontrolled environment. Also the removal of the restriction on circular object movement to obtain a reconstruction refinement will be part of our future work.

Acknowledgments. We would like to thank the reviewers for their helpful and positive comments. This work was supported by the Austrian Research Promotion Agency (FFG) under the project SILHOUETTE (825843).

References

1. Byne, J., Anderson, J.: A CAD-based computer vision system. Image and Vision Computing 16(8), 533–539 (1998)
2. Costa, M.S., Shapiro, L.G.: 3D object recognition and pose with relational indexing. Computer Vision and Image Understanding 79(3), 364–407 (2000)
3. Cyr, C., Kimia, B.: 3D object recognition using shape similiarity-based aspect graph. In: ICCV, pp. 254–261 (2001)
4. Fofi, D., Sliwa, T., Voisin, Y.: A comparative survey on invisible structured light. In: SPIE, vol. 5303, pp. 90–98 (May 2004)
5. Gluckman, J., Nayar, S.K.: Catadioptric stereo using planar mirrors. IJCV 44 (2001)
6. Heber, M., Ruether, M., Bischof, H.: Catadioptric multiview pose estimation for robotic pick and place. In: VISAPP, vol. 1, pp. 423–426 (2010)
7. Hernandez, C., Schmitt, F., Cipolla, R.: Silhouette coherence for camera calibration under circular motion. PAMI 29(2), 343–349 (2007)
8. Hu, B., Brown, C., Nelson, R.: Multiple-view 3-D reconstruction using a mirror. Tech. rep., University of Rochester (May 2005)
9. Laurentini, A.: The visual hull concept for silhouette-based image understanding. PAMI 2 (1994)
10. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH 21(4), 163–169 (1987)
11. Nelder, J.A., Mead, R.: A simplex method for function minimization. The Computer Journal 7(4), 308–313 (1965)
12. Reinbacher, C., Ruether, M., Bischof, H.: Pose estimation of known objects by efficient silhouette matching. In: ICPR (2010)
13. Santner, J., Unger, M., Pock, T., Leistner, C., Saffari, A., Bischof, H.: Interactive texture segmentation using random forests and total variation. In: BMVC, London, UK (September 2009)
14. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: CVPR, vol. 1, pp. 195–202 (June 2003)
15. Szeliski, R.: Rapid octree construction from image sequences. In: CVGIP, vol. 58, pp. 23–32 (1993)
16. Ulrich, M., Wiedemann, C., Steger, C.: CAD-based recognition of 3D objects in monocular images. In: ICRA, pp. 1191–1198 (2009)
17. Zhang, Z.: Flexible camera calibration by viewing a plane from unknown orientations. In: ICCV, pp. 666–673 (1999)