Object Recognition and Pose Estimation using Color Cooccurrence Histograms and Geometric Modeling

Staffan Ekvall, a Danica Kragic, b Frank Hoffmann c

a Computational Vision and Active Perception, Royal Institute of Technology, Stockholm, Sweden ekvall@nada.kth.se

b Centre for Autonomous Systems Royal Institute of Technology, Stockholm, Sweden danik@nada.kth.se

c Electrical Engineering and Information Technology University of Dortmund, Dortmund, Germany hoffmann@esr.e-technik.uni-dortmund.de

Abstract

Robust techniques for object recognition and pose estimation are essential for robotic manipulation and object grasping. In this paper, a novel approach for object recognition and pose estimation based on color cooccurrence histograms and geometric model based techniques is presented. The particular problems addressed are: i) robust recognition of objects in natural scenes, ii) estimation of partial pose using an appearance based approach, and iii) complete 6DOF model based pose estimation and tracking.

Our recognition scheme is based on the color cooccurrence histograms embedded in a classical learning framework that facilitates a “winner–takes–all” strategy across different scales. The hypotheses generated in the recognition stage provide the basis to estimate the orientation of the object around the vertical axis. This prior, incomplete pose information is subsequently made precise by a technique that facilitates a geometric model of the object to estimate and continuously track the complete 6DOF pose of the object.

Major contributions of the proposed system are the ability to automatically initiate the tracking process, its robustness and invariance towards scaling and translations and computational efficiency since both recognition and pose estimation rely on the same representation of the object. The performance of the system is evaluated in a domestic environment (living room) with changing lighting and background conditions on a set of everyday objects.

Key words: Object recognition, pose estimation, color cooccurrence histograms, model based tracking

Preprint submitted to Elsevier Science
1 Introduction

Recent progress of service robotics gradually expands the application domain of robotics from manufacturing settings to domestic environment. Since it is impossible to engineer such a dynamic environment, the ability of robust perception is one of the key components of a robotic system. This paper considers the problem of vision based object detection and pose estimation and its application to manipulation and grasping. The process of object manipulation in general involves all aspects of detection/recognition, servoing to the object, alignment and grasping. Each of these processes has typically been considered independently or in relatively simple environments. Given a task at hand together with its constraints, it is, however, possible to provide a system that exhibits robustness in a realistic setting, [22].

An important skill in terms of grasping is the estimation of the three dimensional position and orientation of the object given an image of the scene, [17]. Due to the large number of topologically distinct aspects of an object, many of the techniques based on computing the correspondence between the image and model features, [18], fail to achieve real–time performance. Furthermore, objects are typically highly textured and it is therefore difficult to use simple features like edges or corners to robustly solve the correspondence problem, see Figure 1.

A more natural approach in terms of computational efficiency is the use of appearance based methods, [20], [23], for providing the rough initial estimate followed by a refinement step using, for example, model based methods, [8], [26]. Compared to our previous work, [17], where the same problem was studied, the major contribution is the computational efficiency of the method due to the object representation used for both recognition and pose estimation steps. In addition, the proposed method shows a significant robustness with respect to scaling and translations.

The paper starts with a brief motivation and introduction of related work in Section 2. It continues with an overview of the underlying appearance based method.
for object recognition in Section 3. Section 4 presents the overall model based tracking system used for real–time tracking of object’s pose. An experimental evaluation of object recognition, pose estimation and their integration in a model based tracking system framework is given in Section 5. Finally, Section 6 concludes the paper with a brief discussion of the proposed method and outlines avenues for future research.

2 Motivation and Related Work

A recent study of human visually guided grasps in situations similar to that typically used in robotic visual servoing control, [13], has shown that the human visuo-motor system takes the underlying three dimensional geometric features into account rather than the two dimensional projected image of the target objects to plan and control the required movements. These computations are more complex than those typically carried out in visual servoing systems and permit humans to demonstrate complex manipulation skills across a large range of problems and environments.

We have therefore decided to integrate both appearance based and geometrical methods to solve different steps of a manipulation task. Many similar systems use manual pose initialization where the user establishes the correspondence between the model and object features [8], [12]. Although there are systems for which this step is performed automatically, [11], [19], the proposed approaches are time consuming and not appealing for real–time applications. For a typical household environment, an additional problem complicates the overall task as the objects to be manipulated by the robot are highly textured and therefore not suited for conventional matching approaches based on, for example, line features, [24], [15], [25].

![Fig. 2. The small image shows the training image used to estimate the nearest pose of the object for the current image. Left) the initial pose overlaid on the current image, and right) the final pose obtained by local refinement method.](image-url)
The basic idea of our approach is the following, see Figure 2: after the object is recognized and its position in the image is known, its initial pose is estimated by an appearance based method. An approach similar to ours has been proposed in [20], where three pose parameters are estimated to guide a robotic arm to a predefined pose with respect to the object. Compared to our approach, where the pose is expressed relative to the camera coordinate system, they express the pose relative to the current arm configuration, making the approach unsuitable for manipulators with a different number of degrees of freedom. Their method relies on a Principle Component Analysis (PCA) based representation to estimate the relationship between the object’s pose and current arm configuration. For cases where the input parameters vary significantly, which is commonly the case in natural environments, this method is suboptimal. Therefore, we apply the method proposed in [4] which represents the appearance of objects by means of color cooccurrence histograms (CCH).

Compared to the system proposed in [26], where the network is entirely trained on simulated images, our method learns from real images. As pointed out in [26], the illumination conditions (as well as the background) strongly affect the performance of their system and these can not be easily obtained with simulated images. In addition, the idea of projecting just the wire-frame model to obtain training images can not be employed in our case due to the objects’ texture.

The system proposed in [24], employs a feature based approach where lines, corners and circles are matched to provide the initial pose estimate. However, this initialization approach is not applicable in our case, since, due to the geometry and textural properties, these features are difficult to detect reliably.

3 Color Cooccurrence Histograms

A color histogram of an object is a compact representation of its appearance, [4]. The estimation of histograms is a fast and computationally efficient process as it does not rely on the explicit reconstruction and matching of geometric features, such as lines or corners. “Regular” color histograms do not preserve geometric structures and objects with similar texture but different shape may therefore result in similar color histograms representations. $X$- and $Y$-color histograms preserve some geometric information, as the bins represent color frequencies along individual rows and columns. In [4], it has been shown that color cooccurrence histograms (CCHs) successfully preserve relations between pixels by representing the frequency of color pixel pairs instead of individual pixels.

The number of individual bins in a CCH is much larger than in the case of regular histograms as it grows with the square number of colors. Therefore, it is necessary to reduce the color set to only contain the most representative colors of the object.
In our work, the optimal color scheme for an object is determined by K-means clustering, [9]. Before the CCH is estimated, pixels are segmented according to color as they get assigned to the nearest cluster in color space.

The similarity between two normalized CCHs is computed as:

\[
\mu(h_1, h_2) = \sum_{n=1}^{N} \min(h_1[n], h_2[n])
\]  

where \( h_i[n] \) denotes the frequency of color pixel pairs in bin \( n \) for image \( i \). The larger the value of \( \mu(h_1, h_2) \), the better the match between the CCHs.

The geometrical relations between pixel pairs can be represented in a number of ways. For example, using both angle and distance based CCHs not only stores the color of the pixel pairs, but also the orientation and length of the vector connecting the two pixels. The drawback of using both angle and distance CCHs is that the representation is no longer rotation and scale invariant. Experimental evaluation will show that, for our application, pure CCHs achieve better performance in terms of pose estimation and computational speed than the CCHs augmented by angle and distance information.

### 4 Model Based Tracking System

Computing all six parameters for objects with complex textural properties has proven to be a difficult problem. For our purposes where a service robot operates in a domestic environment, we assume that the objects to be grasped are placed on a planar, horizontally oriented surface, such as a table or shelf. Therefore, our appearance based approach only estimates a single rotational parameter – namely the objects’ rotation around the vertical axis. Since the robot is equipped with a stereo vision system, a rough estimate of the translational pose parameters is easily obtained. If the stereo system is not available, a model based approach can be facilitated to retrieve the complete pose of the object. This has been demonstrated in our previous work, [17].

Our approach combines the accuracy of geometry based methods with the robustness of appearance-based methods in a synergistic fashion where the key idea of the integrated algorithm is to obtain the initial pose estimate using the appearance-based method. This estimate allows it to project features of the object model onto the image. These projected features provide sufficient prior information to initialize the local search and matching of corresponding features in the image. The integrated approach reduces the global correspondence problem to a local tracking problem.
A typical model based tracking system usually involves the following steps: detection, matching, pose estimation, update and prediction, see Figure 3. The input to the algorithm is a wire-frame model of the object. The main loop starts with a prediction step where the state of the object is predicted by means of the current pose (velocity, acceleration) estimate and a motion model. The visible parts of the object are then projected into the image (projection and rendering step). After the features are detected, they are matched to the projected ones and used to estimate the new pose of the object. Finally, the calculated pose is input to the update step.

4.1 Initialization - Object Recognition

The image is first scanned and a matching vote $\mu(h_{\text{object}}, h_{\text{window}})$ is estimated indicating the likelihood that the window contains the object. Once the entire image has been searched through, a vote matrix provides a hypothesis of the object’s location. Figure 4 shows a typical experimental scene and the corresponding vote matrix. In this case, the package of rice, roughly centered in the image, is being searched for. The vote matrix reveals a strong response in the vicinity of the object’s position (black colored square). Several smaller responses occur near the raisin box and books, which contain similar colors.
The local maxima in the vote matrix serve as starting points to initiate the identification of candidate windows. Each window is iteratively expanded by adjacent rows or columns, as long as the new cells give sufficient support for the object. The expansion process stops when the ratio between the average vote in the border cells and the local maxima vote becomes falls short of the threshold $\Phi$. In principle, the optimal threshold value $\Phi$ depends on the object's color distribution and texture. If the threshold is too high, parts of the object may be omitted. If the threshold is too low, the window contains too much background that reduces the signal to noise ratio in the subsequent image processing steps. An experimental evaluation of different threshold values showed that our algorithm achieves similar performance for a range of $\Phi \in [0.3, 0.6]$, as shown in Figure 12. Finally, a fine search at two-pixel resolution calculates additional, improved CCHs for the candidate windows. Details are provided in Section 5.

4.1.1 Rotation Estimation

Once the object has been segmented from the image, its rotation around the vertical axis is estimated. The similarity in appearance of two poses $i$ and $j$ is calculated according to equation 1. Figure 5 shows the dependency between the match value $\mu(i, j)$ and angular separation in object pose $|\alpha(i) - \alpha(j)|$. To improve the robustness, the hypotheses are first weighted by a Gaussian. If the $i$-th training image with a known angle $\alpha_i$ matches the segmented image of unknown pose to a degree $\mu_i$, the likelihood $P(\beta)$ of the object angle rotation $\beta$ is calculated as:

$$P(\beta) = \frac{\sum_{i=0}^{N} \mu_i \cdot g(\beta, \alpha_i)}{\sum_{i=0}^{N} g(\beta, \alpha_i)}$$

(2)

The Gaussian kernel function

$$g(\beta, \alpha) = \frac{1}{\sigma (2\pi)^{1/2}} e^{-\frac{(\beta - \alpha)^2}{2\sigma^2}}$$

(3)

captures the degree to which the vote $\mu_i$ of a training image contributes to $P(\beta)$ based on the distance $\beta - \alpha_i$. The maximum of $P(\beta)$ emerges in vicinity of training images with high match values $\mu$. For the example shown in Figure 5, the match values $\mu_i$ of training images are clearly correlated with the object’s angle of rotation $\alpha_i$. The distribution has a global maximum at $-39\,\text{deg}$, and a second local maximum at $180\,\text{deg}$. The two minima occur at $100\,\text{deg}$. The algorithm estimates the rotation angle of $-39\,\text{deg}$ at the global maximum which is a fairly accurate estimate of the true rotation angle of $-37\,\text{deg}$. 

4.2 Prediction and Update

The system state vector consists of three parameters describing translation of the target, another three for orientation and an additional six for the velocities:

\[ x = [X, Y, Z, \phi, \psi, \gamma, \dot{X}, \dot{Y}, \dot{Z}, \dot{\phi}, \dot{\psi}, \dot{\gamma}] \]  \hfill (4)

where \( \phi, \psi \) and \( \gamma \) represent roll, pitch and yaw angles \[5\]. The following piecewise constant white acceleration model is considered \(2\):

\[ x_{k+1} = Fx_k + Gv_k, \quad z_k = Hx_k + w_k \]  \hfill (5)

where \( v_k \) is a zero–mean white acceleration sequence, \( w_k \) is the measurement noise and

\[ F = \begin{bmatrix} I_{6 \times 6} & \Delta T I_{6 \times 6} \\ 0 & I_{6 \times 6} \end{bmatrix}, \quad G = \begin{bmatrix} \Delta T^2 I_{6 \times 6} \\ \Delta T I_{6 \times 6} \end{bmatrix}, \quad H = [I_{6 \times 6} \mid 0] \]  \hfill (6)

For the prediction and update, the \( \alpha – \beta \) filter is used:

\[ \hat{x}_{k+1|k} = F_k \hat{x}_k, \quad \hat{z}_{k+1|k} = H \hat{x}_{k+1|k} \]

\[ \hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + W[z_{k+1} - \hat{z}_{k+1|k}] \]  \hfill (7)

Here, the pose of the target is used as measurement rather than image features, as commonly used in the literature, \([7], [11])\). An approach similar to the one presented here was considered in \([25]\). This approach simplifies the structure of the filter which facilitates a computationally more efficient implementation. In particular, the dimension of the matrix \( H \) does not depend on the number of matched features in each frame but it remains constant during the tracking sequence.
Fig. 6. first row) An example of tracking a package of raisins: a fairly textured object against a textured background. The estimated pose of the object is overlaid in white. During this experiment a 6mm lens was used and the object was at a distance of approximately 50cm from the camera, and second row) A moving camera and a static object show the ability of the system to cope with significant depth changes and perspective effects.

4.3 Detection and matching

When a new estimate of the object’s pose is available, the visibility of each edge feature is determined and based on the internal camera parameters a model of the object is projected onto the image plane. For each visible edge, a number of image points is generated along the edge. So called tracking nodes are assigned at regular intervals in image coordinates along the edge direction. The discretization is performed using the Bresenham algorithm, [10]. In the next step, a search is performed for the maximum discontinuity (nearby edge) in the intensity gradient along the normal direction to the edge. The edge normal is approximated with four directions: -45, 0, 45, 90 degrees. In each point along a visible edge, the perpendicular distance to the nearby edge is determined using a one–dimensional search. The search starts at the projected model point and the traversal continues simultaneously in opposite search directions until it encounters the first local maximum. The normal displacements are calculated, and the method proposed in [8], is used. Lie group and Lie algebra formalism provide a the basis for representing the motion of a rigid body and pose estimation. Implementation details can be found in [16]. A few images from a tracking sequence are shown in Figure 6.

5 Experimental Evaluation

The proposed system was experimentally evaluated for: i) object recognition, ii) rotation estimation, and iii) full 6DOF pose estimation and tracking.
5.1 Object Recognition

The performance of the proposed CCHs recognition scheme is evaluated and compared to X-Y-histograms based approach. Five objects are included in the test (see Figure 1): rice and raisins package, soda bottle, mug and cleaner. For training, front and back images for each object are used with the background removed. The sizes of the training images range from 13x36 to 74x81 pixels. The performance of the system is evaluated on ten images of varying size, 150x112 to 483x362 pixels, in a natural setting, see Figure 7. Each image includes all five test objects as well as other objects of similar colors.

In order to reduce the effect of varying illumination, color images are normalized prior to the recognition process. After the normalization, the image is scanned through using a search window of 40x40 pixels in size. The window is shifted such that consecutive windows overlap to at least 50%. The histogram of the candidate window is compared with the object histograms according to Equation 1. Each object is usually represented by two histograms to capture its appearance from different sides (back/front). A single histogram is sufficient if the front and back are similar.

During recognition, a few hypotheses are generated for each object and ranked according to their vote values. The performance of the recognition system is evaluated using the following performance criteria:

1. Localization success (LOC) measures the frequency of correct hypothesis.
2. Window number (WINNR) computes the average rank of successful hypotheses. In a robust recognition algorithm, the object should be included among the highest ranked candidate windows.
3. Window size (WINSZ) compares the size of the bounding window with the size of the entire test image. It measures the amount of the remaining background after segmentation. As this value depends on the size of the object in the window, this parameter is only useful when comparing different recognition schema on a set of identical images.
4. Object integrity (INT) determines what fraction of the object is included in the hypothesis. Object integrity is closely correlated with the segmentation threshold. Intuitively, a small window size and high object integrity are conflicting - better integrity is achieved at the cost of additional background in the candidate widow.

Performance parameters for X-Y- and CCHs based recognition are shown in Table 1. CCHs are clearly superior to X-Y-histograms as indicated by the lower average window number (WINNR). X-Y histograms work well for rice and cleaner objects that contain distinctive colors. The high average window number demonstrates their failure to identify the correct segmentation window on the mug and
soda bottle. The CCHs method reliably segments all objects from the test images. The low average window number shows that in most cases the object is bounded by the highest ranked window. In the remaining cases the object is recognized as the second best hypothesis.

Window size (WINSZ) is smaller in the case of the CCHs which results in a better removal of the background. This advantage comes at the cost of reduced object integrity where, in some cases, only 40% of the object pixels are preserved. To summarize, it can be clearly seen that the recognition algorithm based on CCHs clearly outperforms the scheme based on X-Y-histograms, in terms of robustness as well as background reduction.

The performance is also measured by calculating the vote efficiency $V$ represented by a ratio of the vote response between the strongest hypotheses and the entire image:

$$V = \frac{\sum_{x,y \in \text{Area}_{\text{correct}}} v(x,y)}{\sum_{x,y \in \text{Area}_{\text{total}}} v(x,y)}$$  (8)

By maximizing $V$, we estimate optimal values for the number of colors $N$, the maximum pixel distance $d_{\text{max}}$ and the cluster radius $c_{\text{rad}}$ used for K-means clustering. The optimal values are $N = 50$, $d_{\text{max}} = 9$ pixels and $c_{\text{rad}} = 0.7 \times$ (the average distance to the cluster). The right part of Figure 7 shows how $V$ increases as $d_{\text{max}}$ increases. The superior performance to normal color histograms is obvious when comparing $V$ for $d_{\text{max}} = 0$ (which is a normal color histogram) to $V$ for $d_{\text{max}} = 10$. For this test, the desired region occupied 1.7 % of the total image.

For objects considered here, a rectangular shaped window is a good approximation. However, once the object becomes partially occluded, rectangular segmentation windows become suboptimal. We observed that, for partially occluded object, the ratio between the object and background pixels is significantly lower than in the same scene without occlusion. This effect may be reduced to some extent by using more general, not necessarily rectangular shapes for the segmented region. Basically, the windows could be expanded by adding individual cells, rather than entire rows or columns. In terms of timing, object recognition takes 1.4 s on a Sunblade 100 (500 MHz) of which computing the CCHs was the most time consuming step.

5.2 Rotation Estimation

For training, the correct pose of the object is estimated by manually matching corresponding corner points between the image and a wire–frame model of the object. We have implemented a combination of methods proposed in [6] and [1]. For each training image the complete CCH is computed off–line and stored together with the known rotation of the object. To minimize the noise in the training images, the background is manually removed from the images prior to training. During the ex-
<table>
<thead>
<tr>
<th>Object</th>
<th>LOC</th>
<th>WINNR</th>
<th>WINSZ</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XY CO</td>
<td>XY CO</td>
<td>XY CO</td>
<td>XY CO</td>
</tr>
<tr>
<td>Rice</td>
<td>100 100</td>
<td>1.3</td>
<td>1.2</td>
<td>5.8</td>
</tr>
<tr>
<td>Mug</td>
<td>90 90</td>
<td>8.9</td>
<td>1.0</td>
<td>4.7</td>
</tr>
<tr>
<td>Raisins</td>
<td>100 100</td>
<td>3.3</td>
<td>1.3</td>
<td>9.7</td>
</tr>
<tr>
<td>Bottle</td>
<td>100 100</td>
<td>10.0</td>
<td>1.2</td>
<td>12.0</td>
</tr>
<tr>
<td>Cleaner</td>
<td>100 100</td>
<td>1.1</td>
<td>1.0</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Table 1
Localisation success (LOC), window number (WINNR), window size (WINSZ) and object integrity (INT) for the segmentation scheme using X-Y-histograms (XY) and cooccurrence color histograms (CO).

Experimental evaluation we observed that after about 50 training images no significant improvement in the accuracy is gained. At run time, the CCH of the candidate windows are matched to the stored information to retrieve the rotation \( \alpha \) of the object around the vertical axis. The background is not removed from the test images and CCHs are based on all pixel pairs separated by less than 10 pixels, which roughly amounts to 600k pixel pairs per segmented test image. Figure 8 illustrates how the CCH of a training image changes as the object is rotated by 0, 45 and 90 degrees.

70 images of size 100x100 pixels, with a removed background, are used for training. The method is tested on 30 unmodified, previously unseen images of the same size as the training images. The evaluation of the rotation estimation algorithm is based on manually cropped test images large enough to contain the entire object. As mentioned earlier, the best results, with a mean angular error of 18 deg, are obtained using pure CCHs.

The CCHs corresponding to \( \pm X \) deg are commonly very similar since they are...
basically "mirrored". This results in an ambiguous match value distribution. To deal with this problem, pixel pairs corresponding to positive and negative angles are stored in separate bins, see Figure 9. In addition, the most dominating part of the CCH are pixel pairs with a distance 0. Therefore, these pairs are excluded from histograms during rotation estimation. As a confidence value, $C$ for rotation estimation we use the ratio between the magnitude of the two largest matching values

$$C = \frac{\mu_{\text{max}} - \mu_{\text{avg}}}{\mu_{2\text{ndmax}} - \mu_{\text{avg}}} \quad (9)$$

We experimentally determined the optimal values for the number of color clusters $N$ and the width of the Gaussian kernel $\sigma$ by means of cross-validation. For the rice packet, the optimal values are $N = 50$ and $\sigma = 5$. The results for the winner-take-all approach with $\sigma = 0$ are inferior compared to applying convolution.

5.3 Object Recognition and Rotation Estimation

Our object recognition and rotation estimation algorithms are tested in combination where the strongest hypothesis from the former serves as input to the latter. The
results improved significantly compared to the manually cropped test images. The same set of 70 images is used for training and 30 images of realistic scenes, were use for testing. The sizes of the test images were 320x200 pixels.

As an example, the two narrow surfaces of the rice object are easily confused as they appear almost identical, except for a small patch of letters on one of the sides. The right part of Figure 10 shows segmented images of the rice package taken from opposite directions. As a result of this confusion, the match value graph is bimodal, as seen in Figure 10. In our experiments, the algorithm successfully estimates the angle in all cases in spite of this problem. However, a small fraction of noise is enough to make the algorithm select the other alternative. For the purpose of grasping such a symmetric object, however, it is irrelevant whether it is rotated $-90\,\text{deg}$ or $+90\,\text{deg}$.

![Fig. 10. Center: The appearance of rice package rotated $-90\,\text{deg}$ is very similar to the appearance when it is rotated $+90\,\text{deg}$. This results in a bimodal match value graph (left). An example of a match value graph in an unambiguous case is shown to the right.](image)

The average angular error is 6 $\text{deg}$ which is quite remarkable considering that the angles computed by means of manual feature matching already carry an uncertainty of about 5 $\text{deg}$. The explanation for the improvement over the results in Section 5.2
is that the segmentation algorithm efficiently extracts object pixels. Therefore, the images after segmentation contain less background noise, compared to the manually cropped images.

In our application, the main purpose of the appearance based method is to robustly provide a pose estimate that is accurate enough for initialization of corresponding features in the tracking based scheme. The feature based tracking method tolerates angular errors in the initial pose of up to 25 – 30 deg. As shown in the angular error histogram in Figure 11, all of the 30 test cases meet this requirement.

We also tested the robustness of the pose estimation with respect to changes in scale, camera angle and noise level. The camera tilt angle is varied between 0 and about 30 deg between test- and training images, which this time contained the raisins package instead of the rice package. The average angular error increases to 17 deg. Thus, it can be concluded that the algorithm is robust with respect to reasonable changes in the camera perspective effects.

We further evaluated the robustness with respect to changes in scale for a range [0.5 – 2.0]. As shown in Figure 11, the angular error remains below 20 deg over a range [0.8 – 2]. In our application, the table area that can be reached by the manipulator is fairly limited, such that the distance to the object to be grasped does not vary significantly. For applications in which the distance between object and camera is more uncertain, it may become necessary to perform additional training at wider range of scales and orientations.

Random pixels are added to the test images in order to test the robustness towards noise. In Figure 12, the impact of image noise on the mean angular error is shown. Noise levels above 40% cause a considerable decrease in performance. This is easily explained by the fact that the information stored in a CCH is already corrupted if one of the two pixels is effected by noise or occlusion. At a noise level of 40% per pixel, effectively only 36% of the pixel pairs remain intact. This observation underlines the need for proper object segmentation prior to the pose estimation step. We note here that an angular error of 25 – 30 deg is still sufficiently accurate for proper initialization of the model based pose estimator. In terms of timing, the execution time for the pose estimation step on a Sunblade 100 (500 MHz) was 0.3 seconds.

5.4 Full 6DOF Pose Estimation

In the integrate scheme, object recognition and rotation estimation serve as the initial values for the model based pose estimation and tracking algorithm. The distance of the object from the camera, Z is estimated according to the ratio between the height of the segmented window and the height of the object (which is known from the model) together with the camera parameters. Similarly, X and Y are estimated from the window position in the image. The rotation of the object around
the vertical axis is obtained from the rotation recognition step, while the remaining two angles are initialized to zero.

Figure 13 shows a few examples of processing steps in the integrated scheme. With the incomplete pose calculated in the recognition (first figure from the left) and orientation estimation step, the initial full pose is estimated (second figure from the left). After that, a local fitting method matches lines in the image with edges of the projected object model. The image obtained after convergence of the tracking scheme is shown on the right. Table 2 contains the pose values before and after the fitting stage. It is important to note, that even under the incorrect initialization of the two other rotation angles as zero, our approach is able to cope with significant deviations from this assumption. This is strongly visible in the last row in Figure 13 where, according to the results reported in Table 2, the angle around camera’s Z-axis is larger than 20 deg.

<table>
<thead>
<tr>
<th>Test NO</th>
<th>$X_{bef}$</th>
<th>$X_{aft}$</th>
<th>$Y_{bef}$</th>
<th>$Y_{aft}$</th>
<th>$Z_{bef}$</th>
<th>$Z_{aft}$</th>
<th>$\phi_{bef}$</th>
<th>$\phi_{aft}$</th>
<th>$\psi_{bef}$</th>
<th>$\psi_{aft}$</th>
<th>$\gamma_{bef}$</th>
<th>$\gamma_{aft}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>-41</td>
<td>-19</td>
<td>125</td>
<td>147</td>
<td>590</td>
<td>802</td>
<td>45</td>
<td>22</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Test 2</td>
<td>-93</td>
<td>-89</td>
<td>206</td>
<td>265</td>
<td>748</td>
<td>976</td>
<td>40</td>
<td>16</td>
<td>0</td>
<td>-4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Test 3</td>
<td>-72</td>
<td>-66</td>
<td>138</td>
<td>159</td>
<td>587</td>
<td>774</td>
<td>30</td>
<td>16</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Test 4</td>
<td>-112</td>
<td>-110</td>
<td>120</td>
<td>143</td>
<td>584</td>
<td>756</td>
<td>30</td>
<td>13</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>-22</td>
</tr>
</tbody>
</table>

Table 2
Values show object’s pose before and after the fitting stage. Test1-4 represent experiments shown in Figure 13.

Fig. 12. Left) Angular error as a function of image noise, and Right) segmentation threshold $\theta$. 
6 Conclusions

Object recognition and pose estimation are basic prerequisites for robust robotic manipulation and object grasping. In this paper, a novel approach for object recognition and pose estimation based on color cooccurrence histograms and geometric model based techniques have been presented. The particular problems addressed were: i) robust recognition of objects in natural scenes, ii) estimation of partial pose using an appearance based approach, and iii) complete 6DOF model based pose estimation and tracking.

It has been demonstrated that CCHs are computationally efficient for representing the appearance of an object in the context of object recognition and partial pose estimation. Because of their invariance to scaling and translations, the algorithm performs robustly in natural settings. The proposed scheme applies the same representation of the object’s appearance for recognition and pose estimation. In 84% of cases, the recognition scheme correctly identified the object while, for the remaining cases, the object was captured by the second best hypothesis.

On a basis of 70 training, our scheme consistently estimates the object poses of all
30 test images with a maximum angular error of less than 20 deg and an average angular error of 6 deg. The method is sufficiently robust towards variations in camera angle and scale and is partially able to cope with image noise and occlusion. The main drawback of CCHs is the same as for other color-based methods: low robustness towards changes in lighting conditions. Even though normalization of colors reduces the effect of this problem, it is not enough to cope with the demands of a real, dynamic environment. If the light changes are large enough to move the colors out of their color cluster, or even worse, into another color cluster, the CCH approach fails completely. Experiments with moving color clusters, that automatically adjusts to the new lighting conditions, failed to improve the robustness since it requires that the light change is homogeneous. This problem was observed only during rotation estimation. This observation is explained by the fact that CCHs of the object in different poses are much more similar to each other than the CCHs of the object compared to the background. The robustness towards lighting conditions is one of the topics of our future research.

We believe that the major contribution of this work is in the integration of different techniques to obtain real–time, on–line 6DOF pose estimation. This is still one of the few systems that is able to perform automatic initialization of the pose tracking algorithm.

References


