2DOF POSE ESTIMATION OF TEXTURED OBJECTS WITH ANGULAR COLOR COOCCURRENCE HISTOGRAMS

Thomas Nierobisch and Frank Hoffmann
Chair for Control and Systems Engineering, Department of Electrical Engineering and Information Technology
Universität Dortmund, Germany

Keywords: 2DOF pose estimation, color cooccurrence histogram, probabilistic neural network

Abstract: Robust techniques for pose estimation are essential for robotic manipulation and grasping tasks. We present a novel approach for 2DOF pose estimation based on angular color cooccurrence histograms and its application to object grasping. The representation of objects is based on pixel cooccurrence histograms extracted from the color segmented image. The confidence in the pose estimate is predicted by a probabilistic neural network based on the disambiguity of the underlying matchvalue curve. In an experimental evaluation the estimated pose is used as input to the open loop control of a robotic grasp. For more complex manipulation tasks the 2DOF estimate provides the basis for the initialization of a 6DOF geometric based object tracking in real-time.

1 INTRODUCTION

This paper is concerned with vision based 2DOF pose estimation of textured objects based on monocular views. Object pose estimation is an active area of research due to its importance for robotic manipulation and grasping. The literature reports two distinct approaches to solve the pose estimation problem. Model based methods rely on the extraction of specific geometric features in the image such as corners and edges (Shapiro and Stockman, 2001). The extracted features are then compared and related to a known geometric model of the object. Efficient and reliable approaches for model-based pose estimation with known correspondences have been proposed by (Dementhon and Davis, 1992; Nister, 2003). The drawback of this method is the lack of robustness in the extraction of distinguishable features in particular for textured objects. In addition, feature based methods require the solution of the correspondence problem, which becomes inherently more difficult in case of occlusion and undistinguishable features. In contrast, global appearance based methods capture the overall visual appearance of an object (Schiele and Pentland, 1999). Neither do they depend on the extraction of individual features nor do they face the correspondence problem. This paper follows the latter approach for robust and computationally efficient pose estimation of multi-colored, textured objects. (Chang and Krumm, 1999) proposed distance color cooccurrence histograms for object recognition. They emphasize the conservation of geometric information as the major advantage of color cooccurrence histograms compared to regular color histograms. Based on this fundamental idea, (Ekvall et al., 2005) proposed color cooccurrence histograms for object recognition as well as 1DOF pose estimation. The angular extension of color cooccurrence histograms was suggested by (Nierobisch and Hoffmann, 2004) in the context of pose estimation of robot players (AIBO's). In addition the 2DOF pose estimation of objects with minimal texture and only three distinct colors has been successfully demonstrated. The aim of this paper is to investigate the potential of angular color cooccurrence histograms for 2DOF pose estimation of multi-colored, textured objects. Recently (Najafi et al., 2006) introduced a method that combines appearance and geometric object models in order to achieve robust and fast object detection as well as 2DOF pose estimation. Their major contribution is the integration of the known 3D geometry of the object during matching and pose estimation by a statistical analysis of the distribution of feature appearances in the view space. Nonethe-

Nierobisch T. and Hoffmann F. (2007).
2DOF POSE ESTIMATION OF TEXTURED OBJECTS WITH ANGULAR COLOR COOCCURRENCE HISTOGRAMS.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 52-59

less their approach requires a 3D model of the object, which is difficult to generate for objects of complex shape.

This paper is organized as follows: Section II provides an introduction to color cooccurrence histograms with the focus on the angular extension of the representation. Section III explains the significant steps of segmentation and 2DOF rotation estimation based on color cooccurrence histograms. Section IV introduces a method to predict the confidence in the 2DOF rotation estimation. Experimental results for pose estimation of textured objects are presented in Section V and a summary is provided in Section VI.

2 COLOR HISTOGRAMS

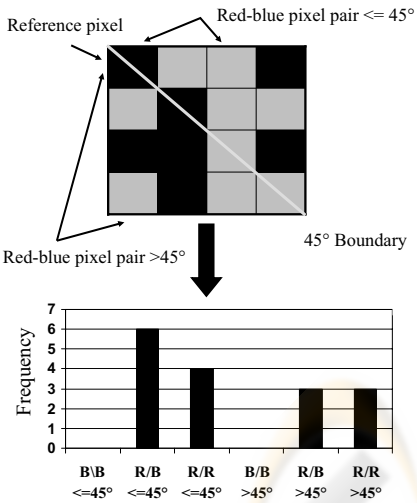

Figure 1: Angular Color Cooccurrence Histogram.

Standard color histograms are often used to capture and abstract the appearance of objects and environments, e.g. for localization tasks (Ulrich and Nourbakhsh, 2000). The drawback of conventional color histograms is that geometrical information about the color distribution is lost in the compression. As a remedy to this detriment (Chang and Krumm, 1999) introduce the color cooccurrence histograms (CCHs), which summarize the geometric distribution of color pixel pairs within the image of an object. An extension of CCHs are angular or distance color cooccurrence histograms (ACCHs or DCCHs, respectively), which contain additional geometric information by either including the orientation of a pixel pair or its distance. A CCH computes pixel pairs in a local environment by starting at a reference. These pixel pairs describe the color information of the reference pixel in conjunction with all other pixels in its local environment. By linearly shifting the reference pixel and its local environment pixel by pixel a geometric color statistics of a region of interest (ROI) is obtained. In order to render the representation independent of scale and size the histogram is normalized. ACCHs augment the geometrical information in comparison to CCHs by additionally storing the orientation of the vector connecting the two pixels. Starting from the reference pixel the angle between the reference frame and a pixel in the local environment is computed and mapped on to a discrete set of angular intervals. Accordingly, DCCHs include statistics on discrete distances between pixels rather than angles. Figure 1 illustrates the method of ACCHs calculation for an image with only two distinct colors. The angular range is discretized into two segments, below or above 45°. Therefore, the histogram consists of six separate bins, namely blue-blue, blue-red and red-red pixel pairs at two distinct angles. The pose estimate relies on the similarity between the stored histograms of known poses with the histogram of the object with unknown pose. Let $h(d,a,b)$ denote the normalized frequency of color pixels with the discrete colors a and b oriented at a discrete angle d. The similarity of two normalized ACCHs is defined as

$$s(h_1,h_2) = \sum_{d=1}^{D}\sum_{a=1}^{C}\sum_{b=1}^{C} \min(\mathrm{h}_1(d,a,b),\mathrm{h}_2(d,a,b)), \quad (1)$$

where h_1 denotes the angle color histogram of the segmented patch in the test image and h_2 is the histogram of a training image stored in the database. D corresponds to the number of angular discretizations and C characterizes the number of distinct colors in the histograms. To obtain a scale invariant similarity measure, termed match value in the remainder of this paper, the histogram counts are normalized by the size of the ROI histogram $\#h_1$,

$$m(h_1,h_2) = \frac{s(h_1,h_2)}{\#h_1}. \quad (2)$$

3 2-DOF ESTIMATION OF OBJECTS WITH TEXTURE

3.1 Experimental Setup

A 5-DOF manipulator and a turntable are used to automatically generate object views at constant distance between camera and object across a view hemisphere. The reference views cover the upper hemisphere, for which due to the limited workspace of the manipulator the elevation is restricted to a range from 45° to

90°. In the following the elevation is denoted by θ and the azimuth by φ. Figure 2 shows the setup for capturing sample images and indicates the view point range. Prior empirical evaluations suggest that a successful open loop grasp requires an accuracy of 10° in θ and 20° in φ for the reference pose. Larger pose estimation errors result in a failure of the open loop grasping controller. Therefore, our performance measure considers angular errors in θ above 10° and in φ above 20° as failures.

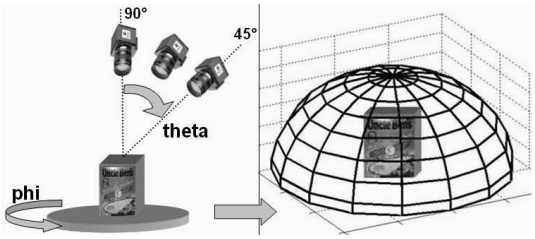

Figure 2: Setup for generating spherical views of the object.

3.2 Object Recognition and 1-D Pose Estimation

In (Ekvall et al., 2005) the authors present an appearance based method for robust object recognition, background segmentation and partial pose estimation based on CCHs. The approach employs a winner-take-all-strategy in which the appearance of an object of unknown pose is compared with a set of training images of known pose. The pose associated with the best matching training image predicts the azimuth orientation of the object around the vertical axis. This prior, incomplete 1DOF pose estimate is subsequently augmented to a complete 6DOF pose by a feature based technique that facilitates a geometric model of the object. In experimental evaluations the average angular estimation error was 6°. Our work is an extension of the previous approach in that it estimates 2DOF spherical poses located on a hemisphere. In addition to the azimuth estimate it also considers the elevation of the camera along the hemisphere. For typical objects with a small top surface the variation of colors along the elevation angle is substantially smaller than for rotations around the vertical axis. Due to this property standard CCHs are unable to capture variations of the object's appearance at large elevation angles. This observation motivates the application of ACCHs for the task of 2DOF pose estimation. Our approach employs the same scheme proposed by (Ekvall et al., 2005) for object recognition and background segmentation to determine the ROI prior to the CCH computation itself. For object recognition the image is first scanned and a matching vote indicates the likelihood that the window contains the object. Once the entire image has been searched, the maximum match provides an hypothesis of the objects location. For background segmentation the best matching window is iteratively expanded by adjacent cells to obtain the final ROI. In a region growing process neighboring cells that bear sufficient resemblance with the object's CCH are added to the ROI.

3.3 2-D Pose Estimation

The purpose of this work is to analyze extended CCHs for the task of 2DOF pose estimation. We assume that the object stands on a planar, horizontally oriented surface and the elevation and azimuth of the eye-in-hand manipulator configuration relative to the object are unknown. From geometric reasoning it is straightforward to identify the type of color cooccurrence histogram (CCH, ACCH or DCCH) which captures the geometric information relevant for 2DOF rotation estimation. CCHs do not contain sufficient information to discriminate between arbitrary poses, as they only count the frequency of pixel pairs but not their relative orientation. In case of a birdseye perspective ($\theta = 90°$) a rotation of the object along the vertical axis (φ) does not alter the frequency of color pairs in the CCH. The same observation applies to DCCHs, as they do not capture the orientation of the vector connecting the two pixels. Obviously, the rotation along the vertical axis does not change the frequency of the color pixels but only their orientation. Therefore ACCHs seem most suitable for 2DOF pose estimation as they are sensitive to variations in appearance that are purely related to the orientation of pixels.
In an experimental evaluation, object views are generated by moving the camera along the vertical axis (φ) in 10° steps from 0° to 360° and along the horizontal axis (θ) from 0° to 180° in 10° steps. To analyze the potential of ACCHs independent of the problem of proper background segmentation, the algorithm is evaluated on a set of views of a textured object in front of a homogeneous background that allows near optimal object segmentation. Compared to the 1DOF pose estimation based on CCH's the 2DOF results are more susceptible to segmentation errors, because the same amount of information is available to extract two degrees of freedom rather than one. Due to the additional angular resolution of the histogram the number of bins in an ACCH is a magnitude larger than for a CCH with the same set of colors. Therefore, the statistics of bin counts in an ACCH deteriorates in comparison to a CCH because the same number of pixel pairs is distributed over a larger number of bins.

Figure 3 shows two match value responses across

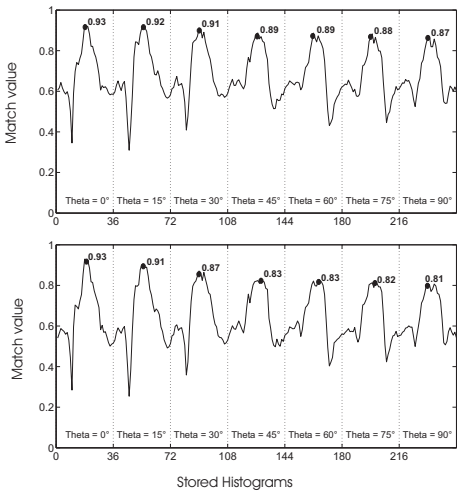

Figure 3: Upper) Match value curve using the local neighborhood for ACCH's Calculation. Lower) Match value curve using a modified approach for ACCH's Calculation.

the training set of image-pose pairs for a test object oriented at a true pose of approximate 183° in φ and 0° in θ direction. The training images are ordered in the sequence $\{[\theta = 0°, \varphi = 0°], [\theta = 0°, \varphi = 10°], ..., [\theta = 0°, \varphi = 350°], [\theta = 15°, \varphi = 0°], ..., [\theta = 30°, \varphi = 0°], ..., [\theta = 90°, \varphi = 350°]\}$.
The match value plot is partitioned into seven slices, each slice corresponding to a full scan along the azimuth φ in $[0 \ldots 360]°$ along seven different elevations of θ in 0°, 15°, 30°, 45°, 60°, 75° and 90°. The shape of the match value response resembles an amplitude modulated signal. The slower modulation corresponds to the horizontal rotation, whereas the faster modulation contains the information about the vertical rotation. The two match value responses correspond to two different ways of computing the ACCH. In the upper match value response pixel pairs originate from a local neighborhood region of the reference pixel. Local color pair statistics are useful to distinguish between multicolored objects with a fair amount of texture. The local statistics results in an ambiguity along the θ rotation because the slow modulation does not discriminate well enough to ensure a robust estimation. In our example the variation in maxima along θ only ranges from 0.93 to 0.87. Pixel pairs counted across a larger separation contribute more information on the object's pose. Nearby pixel pairs, even though useful for object recognition, dilute the information of the object's global appearance as e.g. the likelihood of finding a same colored pixel next to the reference pixel is fairly large. Therefore, the second scheme only counts pixel pairs separated by a minimal distance and ignores pixels in the immediate neighborhood of the reference pixel. As a result the 2DOF appearance of the object reflected through the ACCHs becomes more distinguishable. In this scheme the variation in maxima along θ ranges from 0.93 to 0.81, with a significant decrease in the amplitude of incorrect local maxima. The drawback is that due to the definition of an excluded neighborhood region the scheme is no longer scale-invariant. Therefore, the second approach is only feasible if the relative distance between the camera frame and the object is approximately known in order to properly scale the excluded region.
The test set contains 190 test images with random 2DOF poses that differ from the training set. The mean angular error across the vertical axis is about 10° and 3.8° across the horizontal axis. The ACCHs operate with a resolution of 12 discrete angles and 40 colors. The local environment comprises 20 pixels, but only pixel pairs with a separation of more than 10 pixels contribute to the angular histograms.

4 CONFIDENCE RATING

Our experiments indicate that the major problem for reliable appearance based 1DOF or 2DOF pose estimation in natural scenes is the accuracy of the segmentation in the preprocessing stage. In most cases the 1DOF rotation estimation is fairly robust towards segmentation errors. However if background objects of similar colors are located next to the object the segmentation partially or completely merges the two objects. Incorrect segmentation results in poor performance of the subsequent pose estimation due to the large amount of background noise introduced by the misleading object. In order to detect such incorrect rotation estimates we rate the confidence in an estimate based on the characteristics of the match value response. Estimates that originate from ambiguous match value responses with multiple local maxima of similar magnitude are rejected. A multi-layered feed-forward neural network is trained on match value responses which an expert previously manually classified by visual inspection as either ambiguous or reliable. The match value responses constitute the input vector x_n, based on which the neural network rates the confidence in terms of a probability $h(x_n)$ that the estimate is reliable. The input vector x_n corresponds to the match values of the test image over the set of training images. The training method is similar to the well known backpropagation algorithm, except that in

this case gradient descent minimizes the entropy

$$E_{min} = \min_{w_{ij}} -\sum_{n=1}^{N} d_n \ln(h(x_n)) + (1-d_n)\ln(1-h(x_n)) \quad (3)$$

rather than the squared error (MacKay, 1992). The term $d_n \in \{0,1\}$ denotes the expert reliability classification of the training example x_n. The term w_{ij} denotes the synaptic weights that are subject to optimization. The entropy in Eq. 3 acquires its minimum, if $h(x_n)$ is equal to the relative frequency of training pairs $c(x_n, d_n = 1)/c(x_n, d_n = \{0,1\})$. The classifier reject any rotation estimates with an ambiguous match value response x for which the neural network predicts a confidence lower than $h(x) < 0.8$. For the example shown in the left of figure 4 the rotation estimation fails because the segmentation merges a part of book with similar colors with the object of interest in the foreground. As a result of the poor object segmentation the rotation estimate has an error of about 60°. However, the neural network rejects this rotation estimation due to its low confidence rating of $h(x) = 0.69$ caused by the incorrect segmentation. The corresponding flat match value curve shows two local maxima of similar magnitude. in response to the rejection, the manipulator moves the camera to a different pose in order to capture an image of the object from a better perspective. In the new image shown on the right side of figure 4 the two objects no longer overlap and the segmentation succeeds. The rotation estimation error is less than 20° which is sufficient for the subsequent model based refinement step. The corresponding match value response shows a unique maximum, which the neural network confirms with a high confidence rating $h(x) = 0.98$.

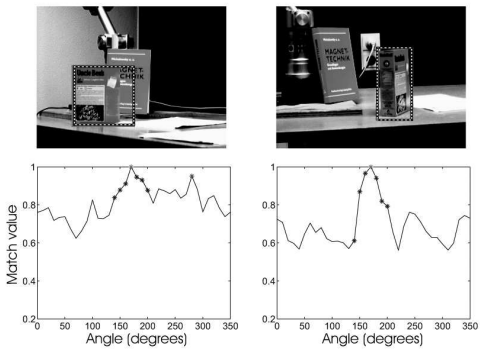

Figure 4: Left top) overlapping objects merged during segmentation. Left bottom) corresponding flat match value response due to incorrect segmentation. Right top) proper segmentation from a different perspective. Right bottom) corresponding match value curve with an unique maximum.

The confidence rating of the 2DOF pose estimate is based on the ambiguity of the match value response. In order to distinguish between reliable and unreliable estimates, a neural network is trained on manually classified match value responses. In order to train the neural network a small subset of features from the match value response that best correlate with the classification has to be selected. From inspection of example responses it turns out, that the distribution and magnitude of global and local maxima are suitable features to predict the confidence.

Figure 5 shows a blue-colored box in the following referred to as object A after being segmented from the background. The pose estimation error in front of the blue background is about 2° in θ and about 46° in φ. The large pose estimation error in φ is caused by the imperfect segmentation of the blue object from the blue background. The error in front of the yellow background that is easier to separate from the object only amounts to 2° in θ and 6° in φ. In the following we analyze the causes for incorrect pose estimates and how to detect potential outliers from the match value response itself, so that these unreliable pose estimates can be rejected beforehand. The corresponding match value responses of object A for the two background scenarios are shown as a 2D-plot (upper graphs) and 1D folded plot (lower graphs) in figure 6. The left top graph shows the ambiguous match value response with several local maxima of similar magnitude caused by poor segmentation in front of the blue background. The noise introduced into the ACCH by the blue background pixels is reflected by the bimodal distribution in φ. In analogy to figure 3 the upper left graph shows the slow modulation corresponding to the changes of θ and the fast modulation corresponding to φ. The stars denote local maxima of the response which magnitude exceeds a threshold of 95 % relative to the global maximum. Additionally also the local maxima along the θ slices are marked by stars in case exceeding 97 % of the absolute maximum. Obviously, several local maxima of similar magnitude in the first slice might correspond to the true object pose in φ.

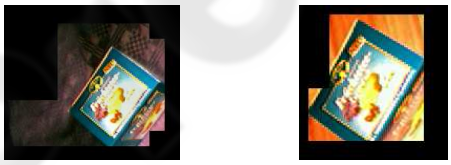

Figure 5: Fruitbox under different background conditions.

The right part of figure 6 shows the 2D and 1D match value responses for the image with proper ob-

ject background segmentation. The 2D match value shows a unique maximum. In the 1D folded representation the local maxima are either in the vicinity of the global maximum or correspond to similar values of φ at different values of θ.

Based on this empirical observation the neural network predicts the confidence based on the following four features extracted from the match value response:

1. R_φ : ratio between the magnitude of the global maximum and the second best match value outside a minimum separation of 20 ° within the same θ slice that contains the global maximum
2. D_φ : separation between the global maximum and the second best match outside the minimum separation within the same θ slice that contains the global maximum
3. R_θ : ratio between the magnitude of the global maximum and the second best match value across all θ slices
4. D_θ : separation between the magnitude of the global maximum and the second best match value across all θ slices

The purpose of the first two features is to detect an ambiguous response in φ, the other two features distinguish between ambiguous and disambiguous responses in θ. Notice, that D_θ is specified in terms of an integer that denotes the number of slices that separate the first from the second maximum. A smooth variation of the match value response with θ implies that the second maximum should occur in the neighboring slice $D_\theta = 1$. Larger values of D_θ in particular in conjunction with a large ratio R_θ indicate a potential ambiguity in the θ estimate. The next section reports experimental results of 2DOF pose estimation and the improvement using the probabilistic confidence rating under real world conditions.

5 EXPERIMENTAL EVALUATION

The experimental evaluation of the proposed methodology in realistic scenarios is based on three test objects of different color and texture with views generated for various backgrounds. The experiments in the previous section assumed an ideal, textureless object with three distinguishable colors in front of a homogeneous background. The purpose of these experiments is to analyze the robustness and accuracy of the pose estimation for daily life objects in a realistic setting. The three test objects are shown in figure 7 and are referred to in the remainder of the text as object A, B and C. The ACCHs operate with a resolution of 10 angles and 40 colors.

Figure 7: Test objects A, B and C.

In the following $\bar{E}$ denotes the mean error of the pose estimate in θ and φ. The training images belong to views at θ angles of 45, 50, 60, 70, 80 and 90°. In φ the sample images are captured in 10° steps. The minimal $\bar{E}$ that is feasible in theory with a winner-takes-all strategy depends on the density of samples, in our case it amounts to 2.5° in φ and 2.4° in θ. The results in table 1 indicate that for all test objects the actual error along θ is close to the optimum. The mean error in φ is significantly lower than the error bounds for successful grasping defined in section III. The experimental results in table 1 demonstrate that under the assumption of near optimal segmentation an error rate of less than 4° in θ and 9° in φ is feasible for all test objects. This error rate is small enough for a successful open loop grasp within the specified error bounds. The percentage of failures is approximately 7%. Table 2 reports the results of the pose

Table 1: 2DOF pose estimation with optimal segmentation.

Objects	Object A	Object B	Object C
$\bar{E}(\varphi)$	7.0°	8.9°	8.5°
$\bar{E}(\theta)$	4.0°	3.3°	3.7°

estimation under different background conditions and the impact of the probabilistic confidence rating on the failure, error and acceptance rate. Background A consists of a wooden material and shows a yellowish textured surface (as shown in the left image in figure 5). The two other backgrounds B and C contain a textured blue and green surface, respectively. The first column specifies the actual object and background, the three following columns describe the mean errors of the pose estimation in θ and φ and the percentage of failures. The next two columns show the mean error $\bar{E}$ for θ and φ for those views that were accepted by the confidence rating based on the match value response. Finally, the percentage of failures and the rate of accepted views (FR) is provided in the last two columns. To verify the methodology under realistic environmental conditions the test set contains 50 images of the three objects taken at random 2DOF positions for the three different backgrounds. Based on the fact that the color distribution of object A contains a large portion of blue colors, segmentation errors with back-

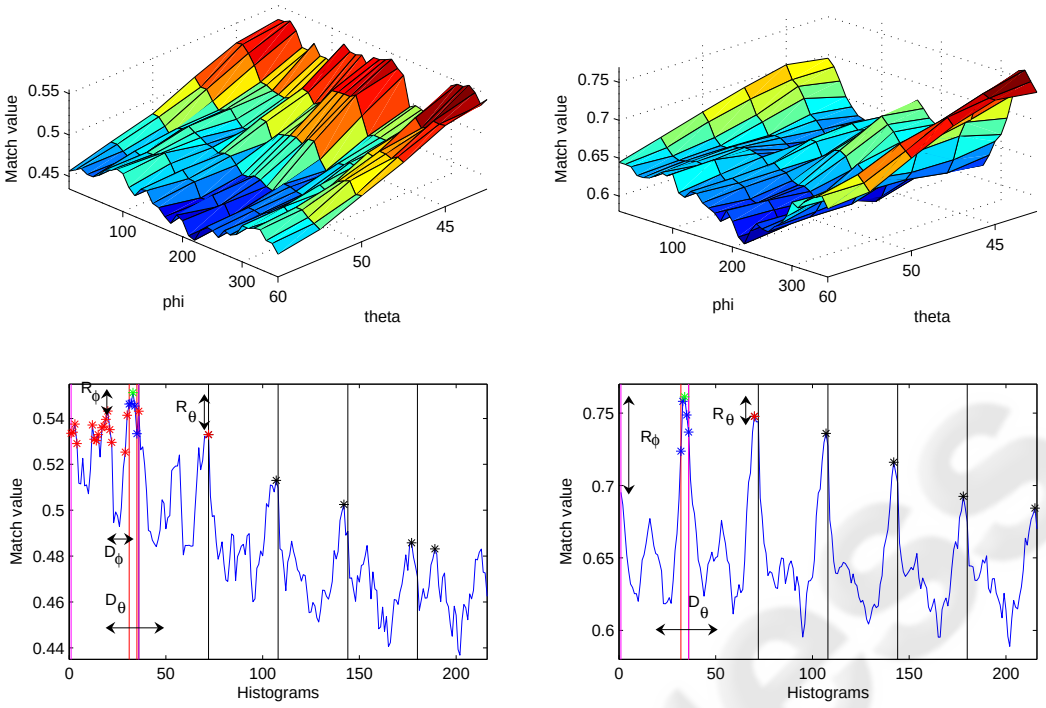

Figure 6: Left top) Ambiguous 2D match value curve based on segmentation noise. Left bottom) According 1D match value response based on segmentation noise. Right top) Unique 2D match value response based on proper segmentation. Right bottom) Corresponding match value response with an unique maximum.

Table 2: 2DOF pose estimation for the three objects under different background conditions.

Object / Background	pose estimation			pose estimation with confidence			
	$E(\varphi)$	$E(\theta)$	Failures	$\bar{E}(\varphi)$	$\bar{E}(\theta)$	Failures	FR
Object A / Backgrd. A	8.0°	11.8°	44%	5.3°	7.3°	27%	30%
Object A / Backgrd. B	22.2°	14.7°	66%	11.2°	9.8°	47%	34%
Object A / Backgrd. C	9.0°	7.2°	20%	4.6°	4.0°	0%	32%
Object B / Backgrd. A	9.1°	15.8°	42%	7.4°	9.2°	19%	42%
Object B / Backgrd. B	10.7°	25.7°	68%	N.A.	N.A.	N.A.	N.A.
Object C / Backgrd. A	12.4°	30.2°	72%	6.1°	19.5°	42%	15%
Object C / Backgrd. B	6.5°	9.1°	28%	4.8°	5.3°	11%	36%

ground B cause a substantial error $\bar{E}$ for θ as well as φ. The results in table 2 demonstrate that for similar object-background colors the color information in a CCH alone provides an insufficient cue for object segmentation and pose estimation. One possible remedy to this problem is to integrate additional cues in the segmentation process. The distance between camera and object or background pixels can be estimated from optical flow or stereo-vision across multiple images taken from slightly different views. It is expected that the segmentation accuracy improves substantially if additional cues are integrated. The objective of the confidence rating is to gain accuracy in the pose estimate, in particular to reduce the number of failures at the cost of rejecting ambiguous object views. In the context of robotic object grasping robust estimation is more important than complete decision making. It is acceptable to reject an ambiguous view and to defer temporarily the grasping process. The manipulator moves the camera to novel viewpoints until the algo-

rithm generates a pose estimate supported with sufficient confidence. For object A the two backgrounds A and C are less problematic in terms of segmentation noise rejection of uncertain poses reduces the mean error as well as the number of failures. In case of background C the neural network is able to exclude all failures, albeit at the cost of rejecting two out three views. Notice, that for test object B in front of background B that coincides with the object color nearly 70% of the original estimates are failures. In this case the neural network ultimately classifies all estimates as unreliable. Acceptable error and failure rates are achieved for test object C in front of background B. The mean estimation error $\bar{E}$ is small enough to allow an open loop grasp for 9 of 10 estimations. In contrast pose estimation on background A fails almost completely with an failure rate of 72% due to the similar colors of the object. Even if only 15 % of the estimates are accepted, the failure rate of 42% is still not acceptable. Instead of an open loop grasp control based on a single image and pose estimate it is more robust to operate in feedback mode by acquiring additional images. A Kalman filter approach fuses observed pose estimates with the known camera motions. The experimental results demonstrate that 2DOF pose estimation based on ACCHs is feasible under the assumption of proper segmentation. The main drawback of the proposed method is the sensitivity with respect to noise and segmentation errors. As a 2DOF pose estimation with ACCHs is substantially more difficult, the approach does not achieve the same level of robustness as in the case of 1DOF pose estimation based on pure CCHs.

6 CONCLUSIONS

In this paper we presented a novel approach for 2DOF pose estimation based on angular cooccurrence histograms. Under the assumption of proper object background segmentation the accuracy of estimated poses is sufficient for object manipulation with a two-finger grasp. The confidence rating of the match value response by the neural network is a suitable means to further improve the robustness of pose estimation at the cost of a reduced recognition rate. The quality of the appearance based segmentation deteriorates substantially in the case of overlapping objects or backgrounds with similar colors. The degradation reflects itself in an ambiguous match value curve detected by the neural network. In a robotic manipulation scenario the camera is moved in order to capture an image of the object from a presumably better perspective. The grasping motion is not executed until a sufficient confidence in the prior pose estimation has been achieved. Our experimental results show that earlier appearance based methods for 1 DOF pose estimation can be extended to a 2DOF pose estimation. However, 2DOF pose estimation based on ACCHs is no longer scale invariant and therefore requires an approximate initial estimate of scale. For our task the reach of the robot arm is limited so that the scale does not vary much across different configurations. Therefore, a single training set of ACCHs captured at an intermediate camera to object range is valid across the entire workspace of the manipulator. An avenue for future research is the integration of appearance based approaches with an image based visual servoing scheme. In image based visual servoing the correspondance problem is prevalent in particular if are only partially visible. To solve the correspondence problem for visual servoing tasks the objects are often labeled with artificial landmarks like color blobs. These approaches are therefore constrained to structured, synthetic environments. To overcome all those limitations visual servoing is established on the entire appearance of an object.

REFERENCES

Chang, P. and Krumm, J. (1999). Object recognition with color cooccurrence histograms. In *CVPR'99,pp. 498-504.*

Dementhon, D. and Davis, L. (1992). Model-based object pose in 25 lines of code. In *ECCV.*

Ekvall, S., Kragic, D., and Hoffmann, F. (2005). Object recognition and pose etimation using color cooccurence histograms and geometric modeling. In *Image and Vision Computing.*

MacKay, D. (1992). The evidence framework applied to classification networks. In *Neural Computation, Vol. 4, 720-736.*

Najafi, H., Genc, Y., and Navab, N. (2006). Fusion of 3d and appearance models for fast object detection and pose estimation. In *Asian Conference on Computer Vision.*

Nierobisch, T. and Hoffmann, F. (2004). Appearance based pose estimation of aibo's. In *International IEEE Conference Mechatronics & Robotics, Proceedings Vol.3, pp. 942-947.*

Nister, D. (2003). An efficient solution of the five-point relative pose problem. In *CVPR.*

Schiele, B. and Pentland, A. (1999). Probabilistic object recognition and localization. In *ICCV'99.*

Shapiro, L. and Stockman, G. (2001). In *Computer Vision.* Prentice Hall.

Ulrich, I. and Nourbakhsh, I. (2000). Appearance-based place recognition for topological localization. In *IEEE ICRA, San Francisco, pp. 1023-1029.*