M. Reinhold, D. Paulus, H. Niemann

Improved Appearance-Based 3-D Object Recognition Using Wavelet Features

appeared in:
Vision, Modeling, and Visualization 2001 (VMV 2001)
21.-23. November, Stuttgart
AKA/IOS Press, Berlin, Amsterdam, p. 473-480

Improved Appearance-Based 3-D Object Recognition Using Wavelet Features

M. Reinhold*, D. Paulus, H. Niemann

Chair for Pattern Recognition, University Erlangen-Nürnberg
Martensstr. 3, D-91058 Erlangen, Germany

Email: reinhold@informatik.uni-erlangen.de

Abstract

In this paper we present an improved appearance-based approach for the localization and classification of 3-D objects in 2-D gray level images. Thereby we calculate local feature vectors by the coefficients of the wavelet multiresolution analysis and model them statistically. Since the appearance of the objects, i. e. also the size of the objects in the image, vary due to out-of-image-plane transformations, the features themselves as well as the region of interest are modelled as function of the external transformations. Further, we present and test different measurements for the recognition of objects that have different sizes. The experiments on a large dataset with more than 40000 images show that the approach is well suited for this recognition task.

1 Introduction

For object recognition there are two main approaches: the approaches that use the results of a segmentation process as features and the approaches that use the image data, i. e. the pixel intensities, directly. In the segmentation process geometric attributes like lines or vertices are detected. Both, themselves [2] and the relationship between them [9] are used as features. But these approaches suffer from two disadvantages: segmentation errors and loss of informations.

Appearance-based approaches avoid these disadvantages. One approach are the so called "multidimensional receptive field histograms" used by [8] that contain the results of local filtering, e. g. by Gaussian derivatives filters. But for modelling 3-D transformations, e. g. a turntable rotation, they need different view-classes, whereby adjacent viewpoints are summarized to one view-class. [4] apply principal component analysis and encode image data as "eigen-images", whereby only one "eigenspace" is necessary for all possible viewpoints. Likewise [1], who model the feature by Gaussian mixtures, needs only one view-class. Admittedly only [4] can classify and localize objects, whereas [8] and [1] can only classify objects, but not estimate its pose.

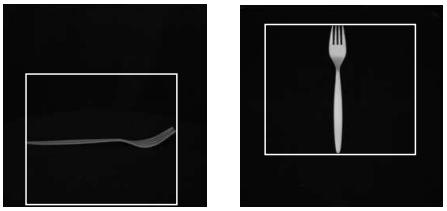

Figure 1: Different viewpoints for a fork. The size of the required fix bounding box is plotted. As one can see the fork takes up only a small part of the bounding box

But all these approaches share the same problem: they employ a fix bounding box that has to be so large that the object lays for all external transformations, i. e. out-of-image-plane transformations, inside the box. If the appearance of the object in the image varies very much for the external transformations, the bounding box encloses very much background and the object takes up only a little part of the bounding box, as one can see in figure 1. But for a reliable recognition only object features and as sparse background features as possible should be considered. The problem can be decreased, but not solved by using view-classes for the different viewpoints.

*This work was funded by the German Science Foundation (DFG) Graduate College "3-D Image Analysis and Synthesis".

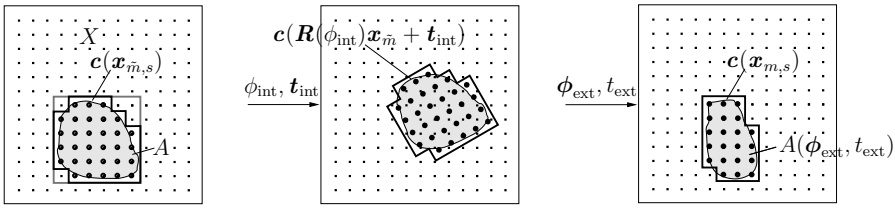

Figure 2: *left:* image covered by the grid for the local feature vectors, the object is enclosed by a tight boundary (black line), the old fix bounding box [5] is plotted in gray; *middle:* the object grid moves with the same internal rotation ϕ_{int} and internal translation $\boldsymbol{t}_{\text{int}}$ as the object; *right:* for the external rotation $\boldsymbol{\phi}_{\text{ext}}$ and the external transformation (scaling) t_{ext} the number of the feature vectors that belong to the bounded region A, i. e. the object, changes

We use and improve the approach of [5]: Local features are derived from the wavelet multiresolution analysis and are modelled statistically by normal distributions. So this approach is robust with respect to noise and to changes in the lightning conditions. For external transformation, the means of the feature vectors are modelled as functions of the external transformations.

In this paper, we additionally model the size of the region of interest - that is the region of the object in the image - as a function of the external transformations, so that this region enclose the object tightly for the respective transformation. The parameter of these functions are learned automatically in the training. The localization and classification of the objects are performed hierarchically by maximum-likelihood estimations. Thereby we propose different measurements and compare their ability to handle objects of different sizes.

In the following section we describe our improved model and present in section 3 the results of the experiments for the old and the improved model. We will end in section 4 with a summary and a outlook.

2 Object Model

2.1 Features

A grid with the sampling resolution $r_s = 2^s$, whereby s is the index for the scale, is laid over the square image $\boldsymbol{f}$ as one can see in the left image of figure 2. These grid locations will be summarized in the following as $X_s = \{\boldsymbol{x}_{\tilde{m},s}\}_{\tilde{m}=0,\ldots,\tilde{M}-1}, \boldsymbol{x}_{\tilde{m},s} \in \mathbb{R}^2$. On each grid point $\boldsymbol{x}_{\tilde{m},s}$ a local feature vector $\boldsymbol{c}(\boldsymbol{x}_{\tilde{m},s}) = \boldsymbol{c}_{\tilde{m},s}$ is calculated. The advantage of local features is that if only one pixel in the image changes, e. g. by noise or occlusion, only the respective feature vector $\boldsymbol{c}_{\tilde{m},s}$ changes, the other feature vectors will be unchanged.

For the feature extraction we perform s times (respective the chosen sampling resolution r_s) the wavelet multiresolution analysis [3] using Johnston 8-TAP wavelets. The first component of the local feature vector $\boldsymbol{c}_{\tilde{m},s}$ is the logarithmic amount of the low-pass coefficient on the respective position, the second component of the local feature vector is the logarithmic sum of the amounts of the first high-pass coefficients on the respective positions [5]. We improve the feature extraction by using not only the integer values of the multiresolution analysis, but the real values. This meliorates the estimation of the variance of the statistical model in subsection 2.3, and thus it improves the recognition rate. To simplify the notation, we omit the index s for the scale in the following.

2.2 Region of Interest

For the object model we lay a close boundary around the object as one can see in the left image of figure 2. We assume that the local feature vectors inside the bounded region $A \subset X$, in the following written as $C_A = \{\boldsymbol{c}_{A,l}\}_{l=0,\ldots,L-1}$ belong to the object and that the feature vectors outside the bounded region $X \setminus A$, written as $C_{X\setminus A} = \{\boldsymbol{c}_{X\setminus A,k}\}_{k=0,\ldots,K-1}$ with $\boldsymbol{c}_{X\setminus A,k} \in \mathbb{R}^2$, belong to the background. For internal rotations ϕ_{int} and internal translations $\boldsymbol{t}_{\text{int}}$ the position of the bounded region A and the position of the object feature vectors $\boldsymbol{c}_{A,l}$ are transformed with the same transfor-

mation as the object (see the image in the middle of figure 2). Thereby the positions $\boldsymbol{x}_m$ of the object grid are calculated by $\boldsymbol{x}_m = \boldsymbol{R}(\phi_{\text{int}})\boldsymbol{x}_{\tilde{m}} + \boldsymbol{t}_{\text{int}}$, whereby $\boldsymbol{R}(\phi_{\text{int}})$ is the rotation matrix. In the following we refer to the object grid with the index m. Since after the transformation the positions $\boldsymbol{x}_m$ of the feature vectors $\boldsymbol{c}_m$ normally do not coincide with the positions $\boldsymbol{x}_{\tilde{m}}$ of the image grid, an interpolation is necessary for calculating the feature vectors $\boldsymbol{c}_m$. In [5] a interpolation scheme that is similar to a linear interpolation was used, in contrast we apply a bilinear interpolation. Additionally, for the calculation of the geometric transformations we employ no longer the inverse rotation matrix, but the transposed rotation matrix; this improves numerical stability.

In [5] the bounded region A was modelled by a rectangular box that was positioned manually. Because of the rectangular form, even for 2-D transformations, there are feature vectors inside the bounding box, i. e. assigned to the object, that properly belong to the background, see the left image in figure 2. Furthermore the size of the bounding box was fix, so for external transformations there are the same problems as we explained in the introduction.

Therefore we improve this approach: the form of the bounded region A was no longer restricted to a rectangle and we now model its size as variable depending on the external transformations. Additionally, its size and form is calculated automatically during the training of the object.

If we have only internal transformations ϕ_{int} and $\boldsymbol{t}_{\text{int}}$, only one image of the object in front of a background that is darker than the object is necessary for the calculation. For the training of the bounded region A, a local feature vector $\boldsymbol{c}_m$ is assigned to the object, i. e. the bounded region A, if the value of its first component (that is derived by the low-pass value of the multiresolution analysis) is higher than a threshold S_A, else it is assigned to the background:

$$c_{m,1} \begin{cases} < S_A \Rightarrow \boldsymbol{x}_m \in X \setminus A \\ \geq S_A \Rightarrow \boldsymbol{x}_m \in A \end{cases} . \quad (1)$$

For the out-of-image-plane rotations $\boldsymbol{\phi}_{\text{ext}}$ and the scaling t_{ext}, we model the form and size of the bounded region A as a function of the external transformations (see right image in figure 2). For this purpose, we define for each local feature vector $\boldsymbol{c}_m$ a function $\xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$ that assigns the local feature vector $\boldsymbol{c}_m$ dependent on the external transformation to the object or to the background. For the training of these functions $\xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$, images of the object in front of a dark background are necessary from different viewpoints. The decision whether a local feature vector $\boldsymbol{c}_m$ belongs to the object or to the background for a certain external transformation is taken analogly to the 2-D case and the value $\xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$ was set respective the assignment to 1 (object) or 0 (background) for this external transformation:

$$c_{m,1}(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \begin{cases} < S_A \Rightarrow \xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) = 0 \\ \geq S_A \Rightarrow \xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) = 1 \end{cases} . \quad (2)$$

For reducing the data-size and also handling viewpoints between the trained viewpoints, we model the single functions ξ_m as continuous and approximate them by a sum of weighted basis functions v_r

$$\xi_m = \sum_{r=0}^{L_\xi - 1} a_{m,r}\, v_r \quad . \quad (3)$$

Thereby we use polynomials for the basis functions v_r. The coefficients $a_{m,r}$ are learned in the training by minimizing the quadratic error between the values of training and the approximated values of $\xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$.

During the recognition, the size and form of the bounded region A is calculated by these trained functions $\xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$ and the local feature vectors $\boldsymbol{c}_m$ are respectively assigned to the background or to the object:

$$\xi_m(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}}) \begin{cases} < S_\xi \Rightarrow \boldsymbol{c}_m \in C_{X \setminus A} \\ \geq S_\xi \Rightarrow \boldsymbol{c}_m \in C_A \end{cases} \quad (4)$$

with $0 < S_\xi < 1$. Thereby the threshold S_ξ determines, how tightly the bounded region A enclose the object. Note that this means no segmentation during the recognition, because the trained the size of the bounded region A is used.

2.3 Statistical Model

We apply a statistical model and interpret the local feature vectors $\boldsymbol{c}_m$ as random variables. Thereby the randomness among others is the noise of the image sampling process and changes in the lightning conditions. Assuming that the object features C_A are statistically independent from the background features $C_{X \setminus A}$, for the object model we

only need to consider the object features C_A inside the bounded region A. Because of the varying size of the bounded region A, it depends on the external transformations, which local feature vectors $\boldsymbol{c}_m$ belong to C_A.

So, the object can be described by the density function $p(C_A|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t})$, depending on the learned model parameter set $\boldsymbol{B}$, the rotation $\boldsymbol{\phi} = (\phi_{\text{int}}, \boldsymbol{\phi}_{\text{ext}})^T$ and translation $\boldsymbol{t} = (\boldsymbol{t}_{\text{int}}, t_{\text{ext}})^T$. We suppose that the single object feature vectors $\boldsymbol{c}_{A,l}$ are statistical independent; in [5] also a row dependency was modelled, but it gave worse results, and therefore we take it no longer into account. Further we assume that the features are normally distributed. So the density function can be written as

$$p(C_A|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t}) = \prod_{\boldsymbol{x}_m \in A} p(\boldsymbol{c}_m|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \boldsymbol{\phi}, \boldsymbol{t}), \tag{5}$$

whereby $\boldsymbol{\mu}_m$ is the trained mean vector and $\boldsymbol{\Sigma}_m$ the trained covariance matrix of the m-th local feature vector. For the supposition that the single components of the feature vectors $\boldsymbol{c}_m$ are statistical independence, $\boldsymbol{\Sigma}_m$ is a diagonal matrix.

Since the appearance of the object varies for the external transformations, the components $\mu_{m,n}$ of the single mean vectors $\boldsymbol{\mu}_m$ depend on the external transformations $\mu_{m,n} = \mu_{m,n}(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})$. We model $\mu_{m,n}$ as continuous functions of the external transformations and approximate them by a weighted sum of basis functions

$$\mu_{m,n} = \sum_{q=0}^{L_\mu - 1} b_{m,n,q}\, v_q \tag{6}$$

using polynomials and learn the coefficients in the training as for (3).

2.4 Localization and Classification

In [5] a maximum likelihood estimation over all possible transformations is performed for the localization:

$$(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{t}}) = \underset{(\boldsymbol{\phi}, \boldsymbol{t})}{\operatorname{argmax}}\, p(C_A|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t}) \quad . \tag{7}$$

For the classification an additional maximum likelihood estimation over all possible object classes is conducted:

$$\hat{\kappa} = \underset{\kappa}{\operatorname{argmax}} \underset{(\boldsymbol{\phi}, \boldsymbol{t})}{\operatorname{argmax}}\, p(C_{A,k}|\boldsymbol{B}_\kappa, \boldsymbol{\phi}, \boldsymbol{t}) \quad . \tag{8}$$

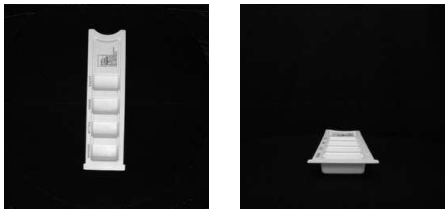

Figure 3: Different viewpoints for a pillbox, the images are 256 pixels in square. In the left image the pillbox takes about 8900 pixels, in the right image about 3800 pixels

This works very well as long we use a fix bounding box so that the number N_A of object feature vectors $\boldsymbol{c}_{A,l}$ is constant for all external transformations and as long all the objects have nearly the same size. But, as one can see in figure 3, the size of an object can vary very much for the external transformations. Besides, the objects in our dataset (see figure 5 below) have different sizes. This causes wrong localizations respective wrong classifications: the density values $p(\boldsymbol{c}_m|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t})$ of the single feature vector $\boldsymbol{c}_m \in C_A$ are normally smaller than 1 even for the right viewpoint and class. Therefore it could happen that a wrong viewpoint respective a wrong class with a distinctly smaller number N_A of object feature vectors than the right viewpoint and class have sometimes a higher density value $p(C_A|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t})$ than the right viewpoint and class.

For this purpose we propose two new approaches for handling the different size of the objects. For the first approach we normalize the density function $p(C_A|\boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{t})$ by the respective number N_A of local object feature vectors C_A for this transformation:

$$(\hat{\kappa}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{t}}) = \underset{\kappa}{\operatorname{argmax}} \underset{(\boldsymbol{\phi}, \boldsymbol{t})}{\operatorname{argmax}} \sqrt[N_{A,\kappa}]{p(C_{A,k}|\boldsymbol{B}_\kappa, \boldsymbol{\phi}, \boldsymbol{t})}. \tag{9}$$

Because of the statistically independence of the local feature vectors $\boldsymbol{c}_{A,l}$ (see (5)) the expression in (9) is the geometric mean of the density values of the single feature vectors $\boldsymbol{c}_{A,l}$. For the localization one respectively for the classification two maximum-likelihood estimations are performed as in (7) and (8).

For the second approach we calculate the ratio of the object density (that is the density for the as-

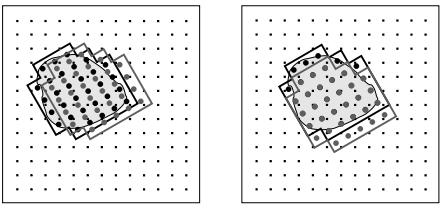

Figure 4: *left:* for each possible internal transformation all the feature vectors have to been interpolated, *right:* if we translate the object grid according to the rotated coordinates axes in steps respective to the chosen resolution r_s, for the same internal rotation only another internal translation the most feature vectors can be reused

sumption that the features C_A belong to the object) and the background density (that is density for the assumption that the same features C_A belong to the background)

$$(\hat{\kappa}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{t}}) = \underset{\kappa}{\operatorname{argmax}} \underset{(\boldsymbol{\phi},\boldsymbol{t})}{\operatorname{argmax}} \frac{p(C_{A,k}|\boldsymbol{B}_\kappa, \boldsymbol{\phi}, \boldsymbol{t})}{p(C_{A,k}|\boldsymbol{B}_0, \boldsymbol{\phi}, \boldsymbol{t})}, \tag{10}$$

whereby $\boldsymbol{B}_0$ are the learned parameters of the background. We use the background model that we have introduced in [6]: the background is modelled as uniform distribution over all possible values of the feature vectors. Because of the statistically independence of the object feature vectors $\boldsymbol{c}_{A,l}$ the background density can be transformed to

$$p(C_A|\boldsymbol{B}_0, \boldsymbol{\phi}, \boldsymbol{t}) = k^{N_{A,\kappa}(\boldsymbol{\phi}_{\text{ext}}, t_{\text{ext}})}, \tag{11}$$

i. e. it depends only from the number $N_{A,\kappa}$ of feature vectors inside the bounded region A. The ratio in (10) is very small (nearly or smaller 1) for a totally wrong position or object and get high for the right position and object.

For accelerating the localization process, it is done hierarchically: it starts on a rough resolution r_{s_0} with a global search followed by a local search. Subsequently, the result of the localization on the rough resolution r_{s_0} is refined on a finer resolution r_{s_1}. For the global search all possible rotations $\boldsymbol{\phi} = (\phi_{\text{int}}, \boldsymbol{\phi}_{\text{ext}})^T$ and translations $\boldsymbol{t} = (\boldsymbol{t}_{\text{int}}, t_{\text{ext}})^T$ are considered and the expressions in (8), (9) respectively (10) are evaluated on discrete points of the n-dimensional transformation space spanned by the possible transformations, with $n \leq 6$.

The global search is computationally very expensive, but it can be strongly speed up by using the redundancy. In [5] an algorithm applying a FFT was employed for this purpose, but this algorithm works only for a fix bounded region A. Therefore we develop a new fast algorithm for the global search. Thereby we use the fact that the interpolation of the grid depends only on the internal transformations ϕ_{int} and $\boldsymbol{t}_{\text{int}}$, whereas the size of the bounded region A and the values of the means $\boldsymbol{\mu}_m$ of the local feature vectors $\boldsymbol{c}_m$ only depends on the external transformations $\boldsymbol{\phi}_{\text{ext}}$ and t_{ext}. Further, for the internal translations $\boldsymbol{t}_{\text{int}}$ we translate the object grid according to the rotated coordinates axes in steps respective to the resolution r_s, so each interpolated feature vector can be used for many internal translations and all external transformations, as visible in the right image of figure 4.

So, we interpolate the required area of the grid for each internal rotation ϕ_{int} only once and store it. Then we calculate the size of the bounded region A and the means $\boldsymbol{\mu}_m$ of the local feature vectors $\boldsymbol{c}_m$ for each external transformations once and combine it with the stored values of the interpolated grid. Thus, the global search over the possible transformations is very fast: in the experiments of section 3 the average computation time for the global search could be reduced from 150 seconds to 3,5 seconds.

3 Experiments and Results

For the experiments we used the 13 objects shown in figure 5 [7]. It is a difficult dataset: Some objects have a similar shape, the appearance and the size of the objects vary very much for the external rotations and especially the cutlery is very small in the image. Besides three different lighting conditions were applied.

We put the objects on a turntable and from each object 3720 gray value images with 256 pixels in square were taken by a camera mounted on a robot arm. Thereby the viewpoints was uniformly distributed over a hemisphere and the angle between two adjacent viewpoints was 3°. Besides the three different lighting was applied so that the lighting is different between adjacent viewpoints. The transformation space consists of 4 dimensions: the external rotation ϕ_{table} with $0^\circ \leq \phi_{table} < 360^\circ$ of the turntable, the external rotation ϕ_{arm} with $0^\circ \leq \phi_{arm} \leq 90^\circ$ of the robot arm, additionally in the experiments we consider the internal translations t_x and t_y. Half of the dataset, i. e. 1860 images

Figure 5: The objects for the experiments: on the one hand office tools like the green and the white-green stapler, the red and the green hole punch, the gray and the red can, on the other hand hospital objects like NaCl-bottle, pillbox, cup with and without saucer and cutlery (fork, knife, spoon)

for each object, was used for the training of the object recognition system, the images for the test were taken from the other half, so training and test set were different.

Although we used a dark background for taking the images, there was left some information in the background like for example the visible edge of the turntable. Since we wanted that for the pose estimation only the object data and no additional background data like the visible edge of the turntable were used, we automatically cut the object out and pasted all the objects in the same homogeneous background with a constant pixel intensity of 15, this is the average value of the original background pixels in the images. For the training of the bounded region A, the objects were pasted in fully black background with a pixel intensity of 0, so we set the threshold $S_A = 0$ (see subsection 2.2). For the training of the object features C_A we laid uniform noise with a pixel intensity $-3 \leq i_n \leq 3$ only over the background (not the object). This is necessary, because for the old model with the fix bounding box [5] there are a lot of background pixels, i. e. also background feature vectors, inside the bounding box A and the statistical model do not work with a synthetical background that has a constant pixel intensity without any noise.

For each approach we performed 466 localization and 154 classification experiments per object. These are altogether 6058 localization and 2002 classification experiments for each approach. In the localization experiments for the rough localization a resolution of $r_{s_0} = 2^3 = 8$ pixels was applied and for the refinement a resolution of $r_{s_1} = 2^2 = 4$ pixels. For the classification experiments we only used the rough resolution r_{s_0}, because this resolution is sufficient for a reliable classification. For the polynomial description of the bounded region A in (3) the means $\boldsymbol{\mu}_m$ in (6) we use polynomials analog to the Taylor decomposition. Admittedly this decomposition is dedicated for external transformations ϕ_{table} and ϕ_{arm} with nearly the same size, whereas the range of the angle ϕ_{table} is four times bigger than the range for the angle ϕ_{arm}. Therefore we need two density functions for each object: one for the region $0^\circ \leq \phi_{table} < 180^\circ$ and one for the

Table 1: Comparison of the error rates of the localization and classification experiments for the old implementation [5] analog eq. (7) and (8), and the new implementation, first with maximizing the geometric mean analog eq. (9), second with maximizing the ratio of the object and the background density analog eq. (10)

	error rate localization experiments			error rate classification experiments		
object	old [5]	mean eq. (9)	ratio eq. (10)	old [5]	mean eq. (9)	ratio eq. (10)
green stapler	12,4%	2,1%	1,7%	33,8%	1,9%	0,0%
white-green stapler	5,2%	0,9%	0,9%	48,1%	0,0%	0,0%
red hole punch	3,0%	0,0%	0,0%	1,3%	0,0%	0,0%
green hole punch	0,2%	0,0%	0,0%	0,0%	0,6%	0,0%
gray can	13,9%	2,4%	4,5%	97,4%	0,0%	0,0%
red can	1,7%	1,7%	1,5%	100,0%	0,6%	0,0%
NaCl-bottle	54,9%	10,1%	10,9%	67,5%	0,0%	0,0%
pillbox	54,3%	23,0%	25,5%	12,3%	0,0%	0,0%
cup with saucer	27,5%	5,4%	4,7%	3,9%	0,0%	0,0%
cup	5,4%	33,5%	9,7%	0,0%	0,0%	0,0%
fork	74,9%	32,6%	37,3%	50,0%	53,2%	39,0%
knife	83,9%	78,3%	71,9%	57,8%	40,3%	8,4%
spoon	78,5%	35,4%	28,8%	92,3%	9,1%	27,3%
total error rate	32,0%	17,3%	15,2%	43,5%	8,1%	5,7%

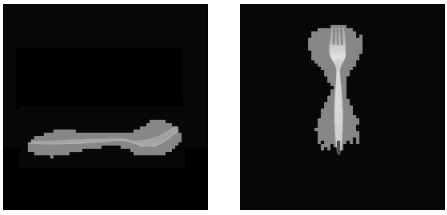

Figure 6: The same viewpoints for the fork as in figure 1. The bounded regions A (in gray) for the threshold $S_\xi = 0,2$ are obviously smaller than the bounding boxes in figure 1. Also, one can see that on the one hand the bounded region A is bigger than necessary and enclose also the pixels of adjacent rotation positions and that on the other hand some pixels at the top and the end of the fork are not enclosed by the bounded region

region $180° \leq \phi_{table} < 360°$. Thereby we used 21 basis functions. We set the threshold $S_\xi = 0,2$ (see eq. (4)). For a higher threshold S_ξ the bounded region will be enclose the object more tightly, but there is also the risk to miss the border of the object. For a lower threshold S_ξ the object will be surely in the bounded region A, but the bounded region A gets bigger. Therefore the threshold $S_\xi = 0,2$ is a compromise. An example for the bounded region A for the threshold $S_\xi = 0,2$ can be seen in figure 6.

In table 1 the results of the localization and classification experiments are presented. A localization was counted as wrong, if the failure for the internal translations t_x and t_y was bigger than 10 pixels or the failure for the external rotations ϕ_{table} and ϕ_{arm} bigger than 15°. Since the appearance of the cup with and without saucer do not change for a rotation $\phi_{table} = 180°$, the localization is also counted as right, if the "failure" for ϕ_{table} is 180°. The localization error rate could be reduced from 32,0% to 17,3% (mean eq. (9)) respectively 15,2% (ratio eq. (10)) by the new approach. Especially for the NaCl-bottle, the pillbox, the fork and the spoon the error rate could be drastically decreased. Only for the knife the error rate is still very high; since it has no distinctive "head" like the fork and the spoon, very often the turntable angle ϕ_{table} is estimated wrong about 180°. Further, it is narrow, therefore even for a observer, it is difficult to gauge the angle ϕ_{arm} of the robot arm with the camera. The two measurements (9) and (10) have nearly

the same average localization rate, for some objects the mean (9) is better, for other the ratio (10). We also tried to use the maximum-likelihood estimation analog to (7) for the localization experiments with the variable bounded region A. But because of the varying size of bounded region A we got an error rate of 36,6%. The computation time for one localizations depends on the size of the object. The average computation time on a Pentium III 800 MHz could reduced from 10 sec for the old approach [5] to 7 sec for the new approach, although for the new approach we additionally had to calculate the size of the bounded region A.

Also the classification rate could be drastically improved: for the old approach the classification error rate was 43,5%, whereas for the new approach it decrease to 8,1% (mean eq. (9)) respectively 5,7% (ratio eq. (10)). Thereby the ratio of the object density and the background density (10) is better than the mean (9): only the cutlery is confused, all the other objects are always recognized. For the maximum-likelihood estimation analog to (8) together with the variable bounded region A, we got an error rate of 74,9%. This shows that for a variable bounded region A the number $N_{A,\kappa}$ of object vectors has to be considered in the measurement. The average computation time for one classification was about 70 sec for the old approach and 50 sec for the new approach.

4 Conclusions and Outlook

In this paper we presented an improved approach for the localization of 3-D objects in 2-D gray-level images. The most important point is that we modelled the size and form of the region of interest (bounded region A) as a function of the external transformations that is learned in the training. Therefore this bounded region A enclose the object very tightly for all viewpoints and is only a little bit larger than the object itself. For the recognition we used the trained form of the bounded region A and so we only considered object features and nearly no background features. We also introduced two new measurements that are suited for the localization and classification of objects of different size and we developed a new fast algorithm for the global search. With this framework we could improve the localization rate from 68,0% to 84,8% and the classification rate from 55,9% to 94,3%, although we have a difficult dataset.

In the future we will extend this approach and combine several object densities and the background density so that we can handle heterogenous background, occlusions and images with multi objects.

References

[1] J. Dahmen, R. Schlüter, and H. Ney, "Discriminative Training of Gaussian Mixtures for Image Object Recognition", in W. Förstner, J. Buhmann, A. Faber, and P. Faber, editors, *21. DAGM Symposium Mustererkennung*, pp. 205-212, Bonn, September 1999, Springer Verlag, Berlin.

[2] J. Hornegger, "*Statistische Modellierung, Klassifikation und Lokalisation von Objekten*", Shaker Verlag, Aachen, 1996.

[3] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 11, Nr. 7, pp. 674-693, Juli 1989.

[4] H. Murase and S. K. Nayar, "Visual Learning and Recognition of 3-D Objects from Appearance", *International Journal of Computer Vision*, Vol. 14, pp. 5-24, 1995.

[5] J. Pösl, "*Erscheinungsbasierte statistische Objekterkennung*", Shaker Verlag, Aachen, 1998.

[6] M. Reinhold, D. Paulus, and H. Niemann, "Appearanced-Based Statistical Object Recognition by Heterogenous Background and Occlusions", *DAGM 2001*, 2001, to appear.

[7] M. Reinhold, Ch. Drexler, and H. Niemann, "Image Database for 3-D Object Recognition", Technical Report LME-TR-2001-02, Chair for Pattern Recognition, 2001.

[8] B. Schiele and J. Crowley, "Object Recognition Using Multidimensional Receptive Field Histograms", *Fourth European Conference on Computer Vision (ECCV)*, pp. 610–619, Cambridge, UK, April 1996, Springer, Heidelberg.

[9] L. G. Shapiro and M. S. Costa, "Appearance-Based 3D Object Recognition", *Object reprensentation in computer vision*, pp. 51-63, Berlin, 1994, Springer-Verlag.