Active Appearance–Based Object Recognition Using Viewpoint Selection

M. Reinhold[1, *], F. Deinzer[1, *], J. Denzler[2], D. Paulus[1], J. Pösl[1]

[1] Chair for Pattern Recognition, University Erlangen-Nürnberg
Martensstr. 3, D–91058 Erlangen, Germany
[2] Computer Science Department, University of Rochester, USA

Email: {reinhold,deinzer}@informatik.uni-erlangen.de

Abstract

In this paper we address the classification of 3–D objects that look similar from several sights and can only be distinguished from some certain viewpoints. For this purpose we combine a statistical appearance-based object recognition approach with an active viewpoint selection mechanism.

For appearance-based object recognition local features are derived from wavelet multiresolution analysis. The recognition process is performed hierarchically in a statistical framework by a maximum likelihood estimation. Based on this result the active viewpoint selection mechanism chooses one further view that allows a reliable classification. Hereby the viewpoint selection mechanism can be trained unsupervised and represents the space of possible viewpoints continuously.

Experimental results show that our approach is well suited for a reliable classification of similar looking objects only by one further view.

1 Introduction

For recognition of 3–D objects in 2–D gray–level images there exist two main approaches in computer vision: based on the results of a segmentation process or directly on the object's appearance. Segmentation operations detect geometric features such as lines or corners. These features as well as the relations between them are used for object recognition [13]. Some authors provide a statistical framework for the geometric features [4, 9]. But all the segmentation approaches suffer from two disadvantages: segmentation errors and loss of information contained in the image caused by the segmentation.

Appearance–based approaches in contrast avoid these disadvantages. They use the image data, i.e. the pixel intensities, directly without a previous segmentation process. The simplest method is correlation of an image with an object template. Another method is the eigenspace approach that was introduced in [6]. Thereby a large number of images are approximatively encoded by a small number of basis images, so-called eigenimages. [11] uses multidimensional receptive field histograms which contain the results of local filtering. We use the appearance–based approach of [8] for object recognition: local features are derived by multi-resolution-analysis and are modelled statistically by parametric density functions. This approach has to be proven to be robust with respect to changes in illumination and to noise.

Approaches that use only a single image for classification and recognition have one disadvantage: if the information in the image is

*This work was partially funded by the German Science Foundation (DFG) Graduate Research Center 3D Image Analysis and Synthesis and under grant SFB 603/TP B2. Only the authors are responsible for the content.

Figure 1: An example for ambiguities between two objects, here two punches, one with a horizontal stripe at the front, one with vertical stripes at the front

not well suited to decide for a certain class and pose, usually an error in recognition occurs or the object is rejected. For example, looking at the image in the middle of Figure 1 it is impossible to decide, which of the punches you see, whether it is the punch with the horizontal stripe (Fig. 1 left) or the vertical stripes at the front (Fig. 1 right). A further view on the front of the punch is required. Therefore we combine our passive appearance–based approach with an active viewpoint selection mechanism.

There are several approaches for viewpoint selection. For example, [7] performs viewpoint selection directly on the extracted outline of an object by a cluster analysis of the trained features. [12] uses a statistical measure - called mutual information - for his active object–recognition system. Performing a statistical classification the most promising viewpoint can be calculated for the complete set of objects in a supervised training step in advance.

Our approach for active viewpoint selection can be trained unsupervised and can handle continuous viewpoints and pose spaces, in contrast to [1], where only discrete positions are possible. We use images of real office objects, in contrast to our previous work presented in [2], where we only examined synthetic images.

In the following section we present our statistical appearance–based object recognition approach. In section 3 we give an overview of our active viewpoint selection mechanism. In section 4 the experimental environment and the results are presented. We conclude with a summary of the results and an outlook to future work in section 5.

2 Appearance–Based Object Recognition

The aim of the presented object–recognition system is the pose estimation and classification of a rigid 3–D object from a 2–D gray–level image. Generally for this task, there are three degrees of freedom for the rotation $\boldsymbol{\phi} = (\phi_x, \phi_y, \phi_z)^T$ and three for the translation $\boldsymbol{t} = (t_x, t_y, t_z)^T$. The transformations can be split into the internal transformations inside the image plane with $\boldsymbol{t}_{\text{int}} = (t_x, t_y)^T$ and $\phi_{\text{int}} = \phi_z$ and the external transformations orthogonal to the image plane with $t_{\text{ext}} = t_z$ and $\boldsymbol{\phi}_{\text{ext}} = (\phi_x, \phi_y)^T$. In contrast to the internal transformations, where the object only changes its position in the image, for the external transformations the object varies its appearance. In this work, the objects were put on a turntable and the camera was fixed. Therefore the distance of the objects to the camera is constant, i.e. $t_z = 0$, and we only have one external rotation ϕ_y, i.e. $\phi_x = 0$.

For the recognition local features are used. Therefore a grid is laid over the quadratic image $\boldsymbol{f} = [f_{ij}]$ with $i, j \in \{0, 1, ..., N-1\}$ as one can see in Figure 2. The distance between the single grid points is $\Delta i = \Delta j = r_S \in \mathbb{R}$, with the sampling resolution r_S for the scale S. In the following these grid locations (i, j) will be summarized as $X_S = \{\boldsymbol{x}_{m,S}\}_{m=0,...,M-1}, \boldsymbol{x}_{m,S} \in \mathbb{R}^2$.

At each grid point $\boldsymbol{x}_{m,s}$ a local feature vector $\boldsymbol{c}_S(\boldsymbol{x}_{m,S})$ that consists of two components is derived from the wavelet multiresolution analysis that is introduced by [5] at the respective scale S. The first component is the logarithmic coefficient of the scaling–function at this position, which is the low–pass coefficient; the second is the logarithmic sum of the amounts of the respective three coefficients of the wavelet–function, which are the high-pass coefficients.

These local wavelet features have three ad-

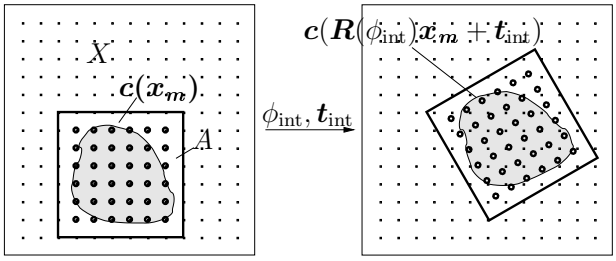

Figure 2: Image covered by a grid for feature extraction, the object grid is moved with same internal transformation as the object

vantages. First, the feature values at certain object locations do not change if the object is translated or rotated in the image plane. Second, if only one single point in the image changes only one single local feature will change, in contrast to this, using global features the whole feature vector will be modified. And third, the multiresolution analysis encourages a hierarchical proceeding that is used for accelerating the localization process.

To simplify the notation, the index S for the scale is omitted in the following. A rectangular bounding box is manually put around the object during the training of the object so that for all external transformations the object will be in the bounding box. We assume that the features $\boldsymbol{c}_A$ at the grid locations inside this box $A \subset X$ belong to our object model and that the features $\boldsymbol{c}_{X\setminus A}$ at the grid points outside this box $X \setminus A$ belong to the background.

As the object is translated and rotated in the image plane, as one can see in the right image of Figure 2, the object grid is moved with the same internal transformations ϕ_{int} and $\boldsymbol{t}_{\text{int}}$ like the object. The new positions of the grid points $\boldsymbol{x}_m{}'$ can be calculated by $\boldsymbol{x}_m{}' = \boldsymbol{R}(\phi_{\text{int}})\boldsymbol{x}_m + \boldsymbol{t}_{\text{int}}$, with $\boldsymbol{R}$ is the rotation matrix and $\boldsymbol{x}_m{}' \in \mathbb{R}^2$. Since the image grid locations and the object grid locations do not match, the values of the features vectors at the transformed position $\boldsymbol{x}_m{}'$ will be calculated by a linear interpolation of the nearest image feature values.

A statistical model is used for object recognition. Hereby the local features $\boldsymbol{c}(\boldsymbol{x}_m)$ are interpreted as random variables. The randomness thereby is, among others, the consequence of noise in the image sampling process and complex changes in the environment (e.g. lighting) conditions.

Assuming that the object features $\boldsymbol{c}_A$ inside the bounding box are independent from the background features $\boldsymbol{c}_{X\setminus A}$ the object model can be described by the density function $p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t})$. It depends on the learned statistical model parameter set $\boldsymbol{B}$, the translation $\boldsymbol{t} = \boldsymbol{t}_{\text{int}}$ and the rotation $\boldsymbol{R} = \boldsymbol{R}(\phi_z, \phi_y)$.

Also, we assume that the feature vectors $\boldsymbol{c}_A$ of the object are normally distributed. Further, we model only a local stochastical dependency, i.e. a feature vector only depends on the former feature vector in its row. This is a reasonable compromise between accuracy and computational complexity. Let $\mathcal{N}(\boldsymbol{c}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the normal distribution, then $\boldsymbol{\mu}$ is the mean vector with concatenated local feature mean vectors $\mu_{m,n}$, where $\mu_{m,n}$ is the mean of the n-th element of the m-th local feature vector, and $\boldsymbol{\Sigma}$ is the covariance matrix with the elements $\sigma_{m,\bar{m},n}$, where $\sigma_{m,\bar{m},n}$ is the covariance between the n-th element of the m-th and the n-th element of the $\bar{m}$-th local feature vector. Because of the row dependency $\boldsymbol{\Sigma}$ is a tridiagonal matrix, all the other components of the matrix are equal to zero. So we obtain for the density function

$$p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t}) = p(\boldsymbol{c}_A|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{R}, \boldsymbol{t}). \quad (1)$$

Since the appearance of the object varies for the external transformation, we model the elements $\mu_{m,n}$ of the mean vector $\boldsymbol{\mu}$ and the elements $\tilde{\sigma}_{m,\bar{m},n}$ of the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ as functions of the external transformation: $\mu_{m,n} = \mu_{m,n}(\phi_y)$; $\tilde{\sigma}_{m,\bar{m},n} = \tilde{\sigma}_{m,\bar{m},n}(\phi_y)$. Assuming these functions are continuous they can be rewritten using a set of basis functions $\{b_r\}_{r=0,\ldots,\infty}$ weighted with appropriate coefficients. We approximate

$$\mu_{m,n} = \sum_{r=0}^{L_\mu - 1} u_{m,n,r} b_r; \quad (2)$$

with L_μ basis functions and

$$\tilde{\sigma}_{m,\bar{m},n} = \sum_{r=0}^{L_\sigma - 1} v_{m,\bar{m},n,r} b_r. \qquad (3)$$

with L_σ basis functions.

The Taylor decomposition shows that the approximation error can be made as small as possible by choosing L_μ resp. L_σ large enough. The value of L_μ resp. L_σ is limited mainly by the size of the training set for estimation. The model parameter $u_{m,n,r}$, $v_{m,\bar{m},n,r}$ are estimated by a maximum likelihood estimation in the training, for further details see [8].

For the localization of the object in an image a maximum likelihood estimation over all possible transformation

$$\underset{(\boldsymbol{R},\boldsymbol{t})}{\operatorname{argmax}}\, p(\boldsymbol{c}_A|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) \qquad (4)$$

is performed.

To speed-up pose estimation, it consists of four steps and is done hierarchically. First, on a rough resolution r_{S_0} the function $p(\boldsymbol{c}_A|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$ is evaluated for a search grid covering the complete possible transformation parameter range. The best g_0 transformation parameters of this global search are improved by a local refinement, implemented by a Downhill–Simplex algorithm [10]. The best result is our pose estimation for the resolution r_{S_0}.

On the finer resolution r_{S_1}, a small, local search grid is laid around the result of the rougher resolution r_{S_0}, and again the g_1 best results of this search grid are improved by a local refinement as before. The best result of the localization on this resolution r_{S_1} is our pose estimation. Hereby the value of the density function $p(\boldsymbol{c}_A|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$ is a measure how good the object model fits the image.

For the classification, at first, for each of the possible K objects the pose is estimated, and afterwards, the decision is reached for the object k with the highest value of the density function. This corresponds to a maximum likelihood estimation over all object classes and all possible poses:

$$k = \underset{\kappa}{\operatorname{argmax}} \underset{(\boldsymbol{R},\boldsymbol{t})}{\operatorname{argmax}}\, p(\boldsymbol{c}_{\boldsymbol{A}_\kappa}|\boldsymbol{B}_\kappa,\boldsymbol{R},\boldsymbol{t}). \qquad (5)$$

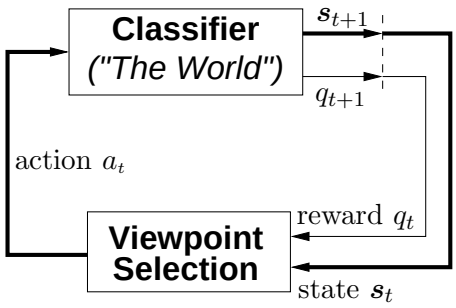

Figure 3: Principles of Reinforcement Learning. A system overview.

This algorithm however has one great disadvantage. If two (or more) objects have the same appearance from a certain viewpoint, as exemplified in section 1, they nearly have the same value of the density function. Also in this case the object–recognition system takes arbitrarily the object class with the highest value of the density function, although it is impossible to classify them in this pose and the classification might be false. Therefore this appearance-based object recognition approach is combined with an active viewpoint selection mechanism which helps to find the right viewpoint for a reliable classification. This active viewpoint selection mechanism will be described in the next section.

3 Viewpoint Selection

The viewpoint selection used for our approach is based on the principles of the *Reinforcement Learning* as illustrated in Figure 3 [14]. The classifier observes at time step t a *state* $\boldsymbol{s}_t$ which represents the estimated class k (cf. eq. (5)) and pose of the object in the real world, which can be written as $\boldsymbol{s} = (k, \phi_y)^T$. Based on this observation the viewpoint selection performs an *action* a_t: the relative movement of the camera around the object ($a \in [0; 360)$). A *reward* q_{t+1} is given for the action a_t which expresses the *significance* "how good can the object be distinguished?" of the resulting view. In contrast to the work in [3] we define the reward as the ratio of the

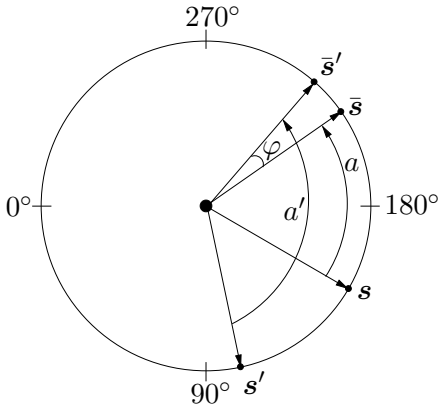

Figure 4: Source states $(\boldsymbol{s}, \boldsymbol{s}')$, executed actions (a, a') and resulting expected destination states $(\bar{\boldsymbol{s}}, \bar{\boldsymbol{s}}')$. The distance $d(\boldsymbol{s}, a, \boldsymbol{s}', a')$ between the expected destination states is denoted as $\sphericalangle(\bar{\boldsymbol{s}}, \bar{\boldsymbol{s}}') = \varphi$.

probabilities of the two best hypothesis

$$q = \frac{\max\limits_{\kappa} \max\limits_{(\boldsymbol{R}_\kappa, \boldsymbol{t}_\kappa)} p(\boldsymbol{c}_A | \boldsymbol{B}_\kappa, \boldsymbol{R}_\kappa, \boldsymbol{t}_\kappa)}{\max\limits_{\lambda, \lambda \neq \kappa} \max\limits_{(\boldsymbol{R}_\lambda, \boldsymbol{t}_\lambda)} p(\boldsymbol{c}_A | \boldsymbol{B}_\lambda, \boldsymbol{R}_\lambda, \boldsymbol{t}_\lambda)} \quad (6)$$

One of the central terms of Reinforcement Learning is the *action–value* function $Q(\boldsymbol{s}, a)$ which describes the quality of performing action a if state $\boldsymbol{s}$ was observed. This quality measure can be written as the expected reward

$$Q(\boldsymbol{s}, a) = E\{q_{t+1} | \boldsymbol{s}_t = \boldsymbol{s}, a_t = a\}. \quad (7)$$

The simplest method (known as Monte Carlo learning [14]) for calculating this expected reward is to average over all rewards which were observed for a state–action pair.

For the training of a Reinforcement Learning system a lot of these action–values $Q(\boldsymbol{s}, a)$ are collected. Usually some ten- or hundred thousands of action–values are necessary for a reliable calculation of the *optimal action*

$$a_{\text{opt}} = \underset{a}{\operatorname{argmax}}\, Q(\boldsymbol{s}, a) \quad (8)$$

for an observed state $\boldsymbol{s}$.

To avoid this time consuming and mostly not very accurate learning principle, an approximation $\hat{Q}(\boldsymbol{s}, a)$ of this action–value function is introduced which is basically the weighted average of all previously collected action–values $Q(\boldsymbol{s}', a')$

$$\hat{Q}(\boldsymbol{s}, a) = \frac{\sum\limits_{(\boldsymbol{s}', a')} K(d(\boldsymbol{s}, a, \boldsymbol{s}', a')) Q(\boldsymbol{s}', a')}{\sum\limits_{(\boldsymbol{s}', a')} K(d(\boldsymbol{s}, a, \boldsymbol{s}', a'))}, \quad (9)$$

where $K(\cdot)$ denotes a suitable kernel function for weighting the distance $d(\boldsymbol{s}, a, \boldsymbol{s}', a')$ of the two state–action pairs $(\boldsymbol{s}, a)$ and $(\boldsymbol{s}', a')$. Small distances — state–action pairs that are "close" to each other — shall be rated high, because their stored information can contribute a lot to the approximation. The calculation of this distance of two state–action pairs is done by measuring the distance between the two *expected destination* states $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{s}}'$. For calculating the expected destination states we currently assume that the pose estimation of the states are correct and that the actions affect the environment in an "optimal" way. For example, in Figure 3 if the estimated pose of $\boldsymbol{s}$ is 150° the action $a = 60°$ results in an expected destination state of $\bar{\boldsymbol{s}} = 210°$. In the case of bad localization results, this definition of expected destination states has to be reconsidered.

We can now define the distance $d(\boldsymbol{s}, a, \boldsymbol{s}', a')$ between the two expected destination states $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{s}}'$ as the angle between the vectors given by the states onto the circle around the object (see Figure 3):

$$d(\boldsymbol{s}, a, \boldsymbol{s}', a') = \sphericalangle(\bar{\boldsymbol{s}}, \bar{\boldsymbol{s}}') \quad (10)$$

with $d(\boldsymbol{s}, a, \boldsymbol{s}', a') \leq 180°$.

This distance leads to our approximation $\hat{Q}(\boldsymbol{s}, a)$ of any action–value (eq. (9)) with the Gaussian kernel function $K(x) = \exp(-x^2/D^2)$. The kernel parameter D specifies how local (small D) or global (large D) the approximation is working. "Local" means, that faraway states have only a slight influence on the approximated action–value with the result of a very detailed approximation. This is useful if you have collected a lot of state–action pairs. On the contrary, a "global" approximation includes data over a wider range of distances and is suitable, if only a few action–values are available.

Figure 5: Two stapler that only differ in the front view, on the left image s1, on the right image s2

Using our approximation $\hat{Q}(\boldsymbol{s}, a)$ we are now able to formulate the search for an optimal action (cf. eq. (8)) as a numerical optimization problem

$$a_{\text{opt}} = \underset{a}{\text{argmax}}\, \hat{Q}(\boldsymbol{s}, a) \tag{11}$$

which can be solved with one of the many well-known numerical techniques. Currently we use an Adaptive Random Search [15] combined with a local simplex for solving eq. (11).

4 Results

Our data set currently consists of four objects, shown in the Figures 1 and 5: a punch marked with a button with a horizontal stripe at the front (in following called object p1), the same punch with a button with vertical stripes at the front (p2), and a stapler with a horizontal stripe (s1) respectively vertical stripes (s2). Therefore, reliable classification is only possible from the front direction, the other directions show ambiguities.

For our experiments the objects were put on a turntable and the camera was fixed, so we have one external rotation ϕ_y for the viewpoint selection. Besides we have an internal translation $\boldsymbol{t}_{\text{int}}$, because during the tests the objects did not exactly stand on the same position on the turntable as during the training. Therefore the search space for pose estimation had three dimensions.

We used Johnston–Wavelets for our appearance-based object recognition. $L_\mu = 10$ basis functions were applied for $\mu_{m,n}$ and $\tilde{\sigma}_{m,\bar{m},n}$ was set as constant, i.e. $L_\sigma = 1$. The localization was performed on two scales, as described in section 2: the roughest had a resolution of $r_{S_0} = 8$ pixels, here $g_0 = 10$ maxima were locally refined, and the finer a resolution of $r_{S_1} = 4$ pixels, here $g_1 = 5$ maxima were locally refined. This choice of the parameter guarantees a high accuracy of the localization. The image size was $256 * 256$ pixels.

object	rec.rate	p1	p2	s1	s2
p1	46%	23	27	0	0
p2	96%	2	48	0	0
s1	70%	0	0	35	15
s2	76%	0	0	11	39

Table 1: Classification results (in percent) and confusion matrix (absolute numbers) for the appearance-based approach without viewpoint selection

For the training of the object recognition 120 images of each object have been taken covering the circle around the object. Hereby two different lighting conditions have been used. The angle between two images is 3° and therefore an object point maximally moves four pixels between two images, which corresponds to the finer resolution $r_{S_1} = 4$.

Afterwards we tested the object recognition on 50 new images of each object, taken from randomly selected viewpoints. The localization of one object requires about 25 seconds on a SGI O2 (R10000, 150 MHz), the mean error for internal translation $\boldsymbol{t}_{\text{int}}$ is 1.2 pixels and for the external rotation ϕ_y is 1.7°. For the classification the pose estimation of all four objects is required, therefore the computing time takes about 100 seconds for an image. The single results of the classification and the confusion matrix are shown in table 1.

The overall recognition rate is 72%. As expected, the object–recognition algorithm confused the two punches as well as the two stapler very often, because three sides of them have the same appearance. Only between 125° and 235° it is possible to distinguish

object	rec.rate	p1	p2	s1	s2
p1	100%	50	0	0	0
p2	100%	0	50	0	0
s1	98%	0	0	49	1
s2	100%	0	0	0	50

Table 2: Classification results (in percent) and confusion matrix (absolute numbers) for the appearance-based approach with active viewpoint selection

them. Hereby the punch 2 has a good recognition rate and punch 1 a bad one, because for a similar appearance the value of density function of punch 2 is nearly always a little bit higher as of punch 1 and therefore the classifier decided for this punch. These results are compared in the following with the active viewpoint selection approach.

For the training the function $Q(\boldsymbol{s}, a)$ has been estimated by performing for each object 30 random movements of the turntable. Being in state $\boldsymbol{s}_t$, i.e. having a class and pose estimate for the object, a random camera movement a_t was chosen. The resulting view was used to classify the object. As a result, the reward was returned, which was stored in $Q(\boldsymbol{s}_t, a_t)$. It is worth mentioning that this is a unsupervised training step. This means also that the system is not told whether or not a classification result is correct.

In Figure 6 you see an example for the action-values and the approximation of $\hat{Q}(\boldsymbol{s}, a)$ of s2 in state $\boldsymbol{s} = (4, 180°)^T$. In this position you see the back of the stapler and it is impossible to take a reliable decision. Therefore the reward for the angles around this position is low and high for a movement in the area from 125° and 235°, where the button with the vertical stripes is visible.

For the test of our viewpoint selection approach the value of the kernel parameter D has been set to 20. For each object 50 experiments have been performed. The turntable has been positioned randomly and an image is taken. Based on the classification and localization result the decision for the next view

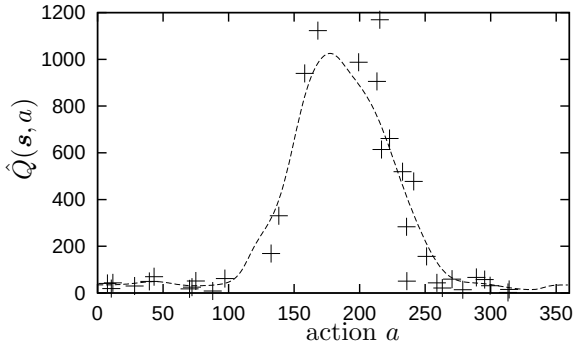

Figure 6: Collected action-values and approximation of $\hat{Q}(\boldsymbol{s}, a)$ for s2 in state $\boldsymbol{s} = (4, 180°)^T$, it is the backside of stapler 2. For the approximation the value of the kernel parameter D has been set to 20.

was made. The next image from the new, calculated viewpoint is taken and used to classify the object. Thus, only one new viewpoint is used in this case.

The classification rates for the four objects using viewpoint selection are shown in table 2. Only one classification failed. The failure was caused by a wrong localization about 180° of s1, and therefore the second viewpoint was ambiguous. So we got an overall classification rate of 99.5% compared to a rate of 72% with a strategy which randomly chooses views.

The computation of one $\hat{Q}(\boldsymbol{s}, a)$ takes about $9 \cdot 10^{-4}$ seconds. The optimization algorithm needs an average of 90 function evaluations of $\hat{Q}(\boldsymbol{s}, a)$ which results in a total time needed for one viewpoint selection of 0.08 seconds.

5 Conclusions

In this paper we have presented a framework for appearance-based object recognition using an active viewpoint selection mechanism. For the recognition local feature derived from the wavelet multiresolution analysis are used, a hierarchical process is performed for accelerating the localization. The statistical appearance-based approach has been proven to be robust with respects to noise and to variations in the lighting conditions.

By combining this approach with an ac-

tive viewpoint selection mechanism the classification results can be improved from 72% to nearly 100%, although the objects can only be distinguished from one direction. Hereby only one action, i.e. one additional view is required. The advantages of our active viewpoint selection mechanism are that it can be trained automatically without user interactions and that the possible viewpoints are continuous in space.

Our future work will concentrate on the acceleration of the object recognition approach, the extension to more external transformations and the handling of more objects.

References

[1] H. Borotschnig and L. Paletta and M. Prantl and A. Pinz, " A Comparison of Probabilistic, Possibilistic and Evidence Theoretic Fusion Schemes for Active Object Recognition", *Computing*, Vol. 62, Number 62, pp. 293–319, 1999.

[2] F. Deinzer and J. Denzler and H. Niemann, "Viewpoint Selection - A Classifier Independent Learning Approach", *IEEE Southwest Symposium on Image Analysis and Interpretation*, Los Alamitos, pp. 209–213, 2000.

[3] F. Deinzer and J. Denzler and H. Niemann, "Classifier Independent Viewpoint Selection for 3–D Object Recognition", to appear in *DAGM 2000*, Kiel, 2000.

[4] J. Hornegger and H. Niemann, "Statistical Learning, Localization, and Identification of Objects", *Proceedings of the 5^{th} International Conference on Computer Vision (ICCV)*, pp. 914-919, Boston, 1995.

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

[6] H. Murase and S. K. Nayar, "Visual Learning and Recognition of 3–D Objects from Appearance", *International Journal of Computer Vision*, Vol. 14, Number 2, pp. 5–24, 1995.

[7] F. Pernus and A. Leonardis and S. Kovacic, "Planning Multiple Views for 3-D Object Recognition and Pose Determination", *Lecture Notes in Computer Science*, Vol. 1296, pp. 424–431, 1997.

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

[9] A. Pope and D. Lowe, "Learning Object Recognition Models from Images", *Early Visual Learning*, eds. Shree Nayar and Tomaso Poggio, Oxford University Press, pp. 67-97, 1996.

[10] W. H. Press and B. P. Flannery and S. A. Teukolsky and W. T. Vetterling, "*Numerical Recipes in C - the Art of Scientific Computation*", Cambridge University Press, New York, 1990.

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

[12] B. Schiele and J. L. Crowley, "Transinformation for Active Object Recognition", *Proceedings of the Sixth International Conference on Computer Vision*, Bombay, India, pp. 249-254, 1998.

[13] L. G. Shapiro and M. S. Costa, "Appearance-Based 3D Object Recognition", *Object Representation in Computer Vision*, Springer-Verlag, Berlin, pp. 51–63, 1994.

[14] R. S. Sutton and A. G. Bart, "*Reinforcement Learning*", A Bradford Book, 1998.

[15] A. Törn and A. Žilinskas, "*Global Optimization*", Lecture Notes in Computer Science, Vol. 350, Springer, Heidelberg, 1987.