Statistical Object Recognition for Multi-Object Scenes with Heterogeneous Background

Marcin Grzegorzek * Kailash N. Pasumarthy * Michael Reinhold Heinrich Niemann

Chair for Pattern Recognition
University of Erlangen-Nuremberg
Martensstr. 3, 91058 Erlangen, Germany

Marcin.Grzegorzek@informatik.uni-erlangen.de

Abstract

In this paper we present a statistical, appearance-based approach for localization and classification of 3-D objects in 2-D gray level images, in which the number of objects in a scene is unknown. First the statistical models of all possible object classes are created separately. The local feature vectors that we use are computed based on wavelet transformation and modeled using a normal distribution. Further, we describe a new approach for the recognition in the case of multi-object scenes. Besides the localization and classification problem, we have to estimate the number of objects in the image. For this purpose we have developed a serial search algorithm with a robust abort criterion. The experiments made on a large sample set with more than 9000 test images show that the approach is well suited for this recognition task.

1. Introduction

The automatic recognition of objects in real environment scenes is becoming more important lately. There exist two main approaches to solve this problem: the model- and appearance-based methods. The model-based algorithms use a segmentation step to extract the features of objects [5], the appearance-based methods determine the feature vectors directly from the image data [8, 3]. Segmentation operations detect geometric features such as lines or corners and use relations between them. But all the segmentation approaches suffer from two disadvantages: segmentation errors, and loss of information contained in the image caused by segmentation.

*This work was partly funded by the German Research Foundation (DFG) Graduate Research Center 3D Image Analysis and Synthesis

We chose the appearance-based approach, because we do not have any information about the shape of the objects, and the segmentation step would not work acceptably. It uses the image data, i.e. the pixel intensities, directly without a previous segmentation process. The simplest method is the correlation of an image with an object template [1]. Another method is the eigenspace approach that was introduced in [7]. There are appearance-based algorithms that use one global feature vector for the whole image (e.g. eigenspace approach), and those that use more local feature vectors (e.g. neural networks [11]). In this work, local feature vectors with two components are applied. They are derived by multi-resolution-analysis [2, 6] and modeled statistically by density functions.

In real world environments it is possible that more than one object from a sample set appears in the scene. This is the reason why we developed a new serial search algorithm for multi-object scenes. The starting point of the algorithm is the approach for appearance-based statistical object recognition by heterogeneous background and occlusions in scenes with individual objects. We use it to search for single objects in the multi-object scene. The serial search algorithm is stopped if our abort criterion is fulfilled. There are some publications about object recognition that do not exclude the case of multi-object scenes, but they are based on other approaches [4].

In section 2, the statistical object model and its components are presented. Section 2.1 describes how we compute the local feature vectors. In section 2.2, the so called region of interest (region of object in the image) is defined. Section 2.3 shows how the statistical parameters are modeled and the object density computed, and section 2.4 describes the separate model for the background. The algorithm for the recognition in the case of multi-object scenes is described in section 3. There we shortly review the well-known like-

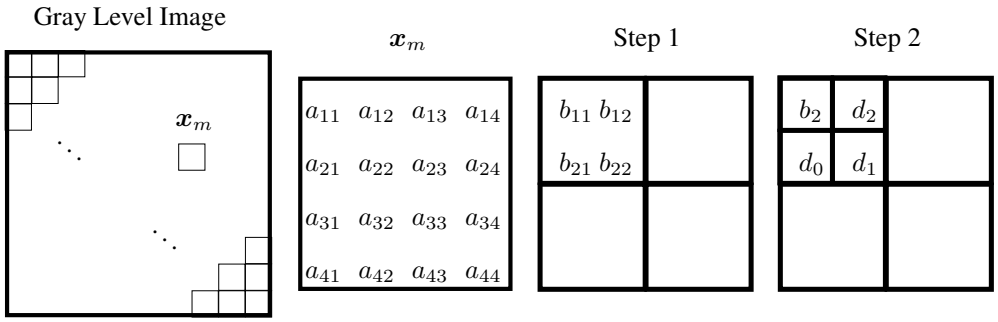

Figure 1. Computation of a feature vector at a grid point $\boldsymbol{x}_m$ with a Haar wavelet (scale $s = -2$). In the first step the local neighborhood of $\boldsymbol{x}_m$ is divided into four squares and low-pass coefficients $b_{i,j} = 0.25 \cdot \sum_{k=0}^{1} \sum_{l=0}^{1} a_{k+2i-1,l+2j-1}$ are computed from the gray values a_{ij}. After the second step b_2 is the low-pass coefficient and is calculated using $0.25 \cdot \sum_{k=1}^{2} \sum_{l=1}^{2} b_{kl}$. The other coefficients result from combinations of low-pass and high-pass filtering ($d_0 = 0.25 \cdot [-(b_{11} + b_{12}) + (b_{21} + b_{22})]$, $d_1 = 0.25 \cdot [-(-b_{11} + b_{12}) + (-b_{21} + b_{22})]$, $d_2 = 0.25 \cdot [(-b_{11} + b_{12}) + (-b_{21} + b_{22})]$).

lihood estimation (section 3.1). In section 3.2 the global assignment function is defined and the serial search algorithm for recognition in multi-object scenes is presented. The experiments and results can be found in section 4. Section 5 will close this paper with a conclusion and a short outlook to further investigations.

2. Statistical Object Model

During the training phase statistical models $\mathcal{M}_\kappa$ of all possible object classes Ω_κ ($\kappa = 1, 2, \ldots, k$) are learned. First we define a set of all possible object classes $\Omega = \{\Omega_1, \ldots, \Omega_\kappa, \ldots, \Omega_k\}$ and take training images of them with a dark background, whereby positions of objects in all images are known. Then we set one of the images for each object class as a reference image. With position of an object in the image $\boldsymbol{f}_i$ we denote the transformation (translation and rotation) that maps the object in the reference image to the object in $\boldsymbol{f}_i$. In the next step the sample set of training images is preprocessed and we get square gray level images. In the following subsections we explain about the components of the model $\mathcal{M}_\kappa$ and the means to get them. The class index κ is omitted in the equations of the current section, because the modeling is identical for all object classes.

2.1 Feature Vectors

In all of the gray level images local feature vectors using a wavelet transformation are computed [2, 6]. A grid with the size $\Delta r = 2^{-s}$, where s is the scale of the wavelet transformation, is laid over each training image. At each grid point a vector $\boldsymbol{c}_m$ with two components is calculated:

$$\boldsymbol{c}_m = \begin{pmatrix} \ln(2^s \, |b_{s,m}|) \\ \ln[2^s \, (|d_{0,s,m}| + |d_{1,s,m}| + |d_{2,s,m}|)] \end{pmatrix} \quad (1)$$

The value $b_{s,m}$ is the low-pass coefficient and $d_{0\ldots2,s,m}$ result from combinations of low-pass and high-pass filtering. An illustration for the computation of a feature vector can be seen in figure 1. Using the local feature vectors has a very important advantage, namely if only one pixel changes its value in the image, e.g. by noise or occlusion, only local feature vectors in a small region vary.

2.2 Region of Interest

Usually, only a part of the image pixels belong to objects. The rest is background. It is not necessary to consider all feature vectors in the whole image. That is why we define for each object class in each training position the region of interest (bounding region). A close non-rectangular boundary is laid around the object. The feature vectors inside this bounding region belong to the object and those outside to the background. The decision is made due to a simple threshold approach, because the training images are taken on a dark background. If there are only internal transformations in the sample set (translations in the image plane, rotation about the orthographic axis to the image plane) the appearance and size of the object do not change. In this case we need only one image to train one object class. The new positions in the object grid $\boldsymbol{x}'_m$ are calculated from the old grid points $\boldsymbol{x}_m$ with following equation:

$$\boldsymbol{x}'_m = R(\Phi_{int})\boldsymbol{x}_m + \boldsymbol{t}_{int} \quad (2)$$

where Φ_{int} denotes the internal rotation angle, $\boldsymbol{t}_{int}$ are the internal translations, and $R(\Phi_{int})$ describes the rotation matrix. For the external transformations (rotations about two orthographic axes in the image plane and scaling) the size and appearance of the object in the image vary. In this case many training images $\boldsymbol{f}_i$ for many different external parameters ($\boldsymbol{\Phi}_{ext,i}, t_{ext,i}$) are needed. For each feature vector $\boldsymbol{c}_m$

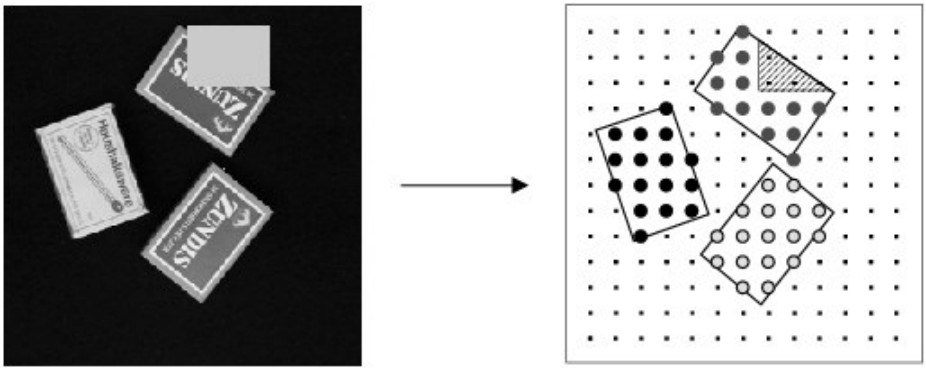

Figure 2. The global assignment function assigns the feature vectors to the objects in the multi-object scene, whereby one feature vector can be assigned maximal to one object in the scene. After a vector was assigned to an object, it is not taken into account in the next steps of the serial search algorithm.

we define a function that assigns it to the object or to the background [10]:

$$\xi_m(\boldsymbol{\Phi}_{ext}, t_{ext}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{c}_m \in O \\ 0 & \text{if} \quad \boldsymbol{c}_m \notin O \end{cases} \tag{3}$$

This function is interpolated using $(\boldsymbol{\Phi}_{ext,i}, t_{ext,i})$, and defined on a continuous domain $(\boldsymbol{\Phi}_{ext}, t_{ext})$. O denotes the bounding region. The internal and external transformations can be written together: $\boldsymbol{\Phi} = (\Phi_{int}, \boldsymbol{\Phi}_{ext})^T$ and $\boldsymbol{t} = (\boldsymbol{t}_{int}, t_{ext})^T$.

2.3 Object Density

After we defined the assignment function, we can separately model the object feature vectors and the background feature vectors. The components of the object feature vectors are statistically modeled using a normal distribution. It means that we compute for each object feature vector $\boldsymbol{c}_m = (c_{m1}, c_{m2})^T$ a corresponding mean value vector $\boldsymbol{\mu}_m = (\mu_{m1}, \mu_{m2})^T$ and a standard deviation vector $\boldsymbol{\sigma}_m = (\sigma_{m1}, \sigma_{m2})^T$. For internal transformations the mean values are constant. Under external transformations the mean values vary and can be written as functions of these transformations $\boldsymbol{\mu}_m = \boldsymbol{\mu}_m(\boldsymbol{\Phi}_{ext}, t_{ext})$. Standard deviations are constant in both cases [8]. We assume that the object feature vectors are statistically independent of the background features. The statistical independence of the single feature vectors and their components is also assumed. The density function for the object features can be described with the following equation:

$$p(C|\boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{t}) = \prod_{\{m|\xi_m=1\}} p(\boldsymbol{c}_m|\boldsymbol{\mu_m}, \boldsymbol{\sigma_m}, \boldsymbol{\Phi}, \boldsymbol{t}) \tag{4}$$

where C denotes the set of all object feature vectors, $\boldsymbol{B}$ comprehends the trained mean vectors and standard deviation vectors, and $(\boldsymbol{\Phi}, \boldsymbol{t})$ are the transformation parameters.

2.4 Background Density

To solve the problem with the heterogeneous background in the recognition phase we introduce also a separate model for background feature vectors. The components of these vectors are modeled using a uniform distribution. Therefore, a priori, nothing has to be known about the background in the recognition phase. All possible backgrounds can be handled by the same model. The background density for a feature vector $\boldsymbol{c}_m \notin O$ is constant for all positions and independent of the transformation parameters $(\boldsymbol{\Phi}, \boldsymbol{t})$.

3. Localization and Classification

After for each Object class Ω_κ the corresponding model $\mathcal{M}_\kappa$ was created, objects can be localized and classified. At the beginning an image is taken, preprocessed and feature vectors are computed with the same method as in the training phase. For the recognition in multi-object scenes a serial search algorithm is applied. This means that we look for the first object in the scene, then for the second, etc. until an abort criterion is fulfilled. The results do not depend on the order in which the objects are searched. In the following section 3.1 we briefly present a modified maximum likelihood estimation algorithm for recognition of individual objects in the scene.

3.1 Modified Maximum Likelihood Estimation

The recognition (localization and classification) with the standard maximum likelihood estimation can be written as follows:

$$(\widehat{\kappa}, \widehat{\boldsymbol{\Phi}}_\kappa, \widehat{\boldsymbol{t}}_\kappa) = \underset{\kappa}{\operatorname{argmax}} \{ \underset{(\boldsymbol{\Phi}, \boldsymbol{t})}{\operatorname{argmax}} p(C_{O_\kappa}|\boldsymbol{B}_\kappa, \boldsymbol{\Phi}, \boldsymbol{t}) \} \tag{5}$$

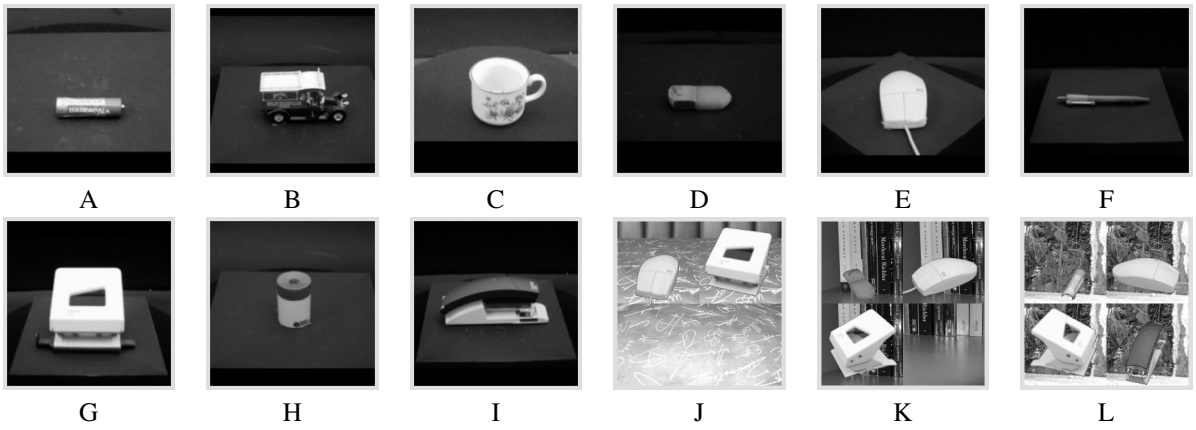

Figure 3. A - battery, B - car, C - cup, D - eraser, E - mouse, F - pen, G - puncher, H - sharpener, I - stapler, J - two object scene, K - three object scene, L - four object scene

For each position hypothesis $(\boldsymbol{\Phi}, \boldsymbol{t})$ we first determine the set of feature vectors C_{O_κ} that have to be taken into account (belonging to the bounding region). $\boldsymbol{B}_\kappa$ contains the statistical parameters of the object class κ. Some feature vectors on the border of the bounding region depend also on the background pixels. Their values are random, because the background is heterogeneous and unknown, and they do not describe the object. Although such feature vectors belong to the region of interest, they should not be taken into account by the evaluation of the object densities. Therefore we defined a new assignment function $\boldsymbol{\zeta}$ that modifies the set of feature vectors C_{O_κ}. The components ζ_m of the function assign the feature vectors $\boldsymbol{c}_m \in C_{O_\kappa}$ to the background or to the object again. $\boldsymbol{\zeta}$ is determined in the maximization process and the whole recognition is described by:

$$(\widehat{\kappa}, \widehat{\boldsymbol{\Phi}}_\kappa, \widehat{\boldsymbol{t}}_\kappa, \widehat{\boldsymbol{\zeta}}_\kappa) = \underset{\kappa}{\operatorname{argmax}}\{\underset{(\boldsymbol{\Phi},\boldsymbol{t},\boldsymbol{\zeta}_\kappa)}{\operatorname{argmax}} p(C_{O_\kappa}|\boldsymbol{B}_\kappa, \boldsymbol{\Phi}, \boldsymbol{t})\} \tag{6}$$

3.2 Serial Search Algorithm

In order to recognize objects in a multi-object scene we introduced a global assignment function. This function assigns the feature vectors in the image to the objects as can be seen in the Figure 2. The feature vector $\boldsymbol{c}_m$ is assigned at most one object in the scene. The already assigned feature vectors are labeled and not used in the next steps of a serial search algorithm. At the beginning the serial search algorithm estimates for all possible object classes Ω_κ the best position hypothesis $({}^1\widehat{\boldsymbol{\phi}}_\kappa, {}^1\widehat{\boldsymbol{t}}_\kappa)$. The prefix "1" in the expression $({}^1\widehat{\boldsymbol{\phi}}_\kappa, {}^1\widehat{\boldsymbol{t}}_\kappa)$ means that the algorithm looks for the first object in the scene. According to the equation (6) the object with the highest probability is recognized. The feature vectors, that belong to the object, are marked in the global assignment function. Subsequently, the search for the second object in the scene is started. The feature vectors mentioned in the global assignment function are not taken into account in the following steps. The searching process is repeated until an abort criterion is fulfilled. The abort criterion tells the system that there are no more valid object hypotheses in the image. For the scene model (heterogeneous background) it is defined as:

$$\frac{N_{C_O,\kappa} - N_{H,\kappa}}{N_{C_O,\kappa}} \begin{cases} < S_p & \Rightarrow \text{ hypothesis not valid} \\ \geq S_p & \Rightarrow \text{ hypothesis valid} \end{cases} \tag{7}$$

$N_{C_O,\kappa}$ is the number of feature vectors from the bounding region that really belong to the object, and $N_{H,\kappa}$ is the number of feature vectors from the bounding region assigned to the background with the function $\boldsymbol{\zeta}_\kappa$. It means that if $S_p \cdot 100\%$ of the object is visible in the image, the position hypothesis is valid. If for the object class Ω_κ there are no valid bounding regions $C_{O,\kappa}$, so this object can not appear in the image. If there are no valid bounding regions for all object classes, there are no more objects in the scene, and the serial search algorithm ends.

4. Experiments and Results

For the training and recognition we used nine objects from the common office environment (figure 3A-I). The position of objects in our sample set is defined with one external rotation. We took 60 images for the training of each object class. The viewpoints are uniformly distributed on a circle and the angle between two adjacent viewpoints amounts $6°$. All training positions can be written as the following sequence: $(0°, 6°, 12°, \ldots, 354°)$. Since the training images were taken on a dark background, the bounding regions (object areas) were trained with a simple threshold approach.

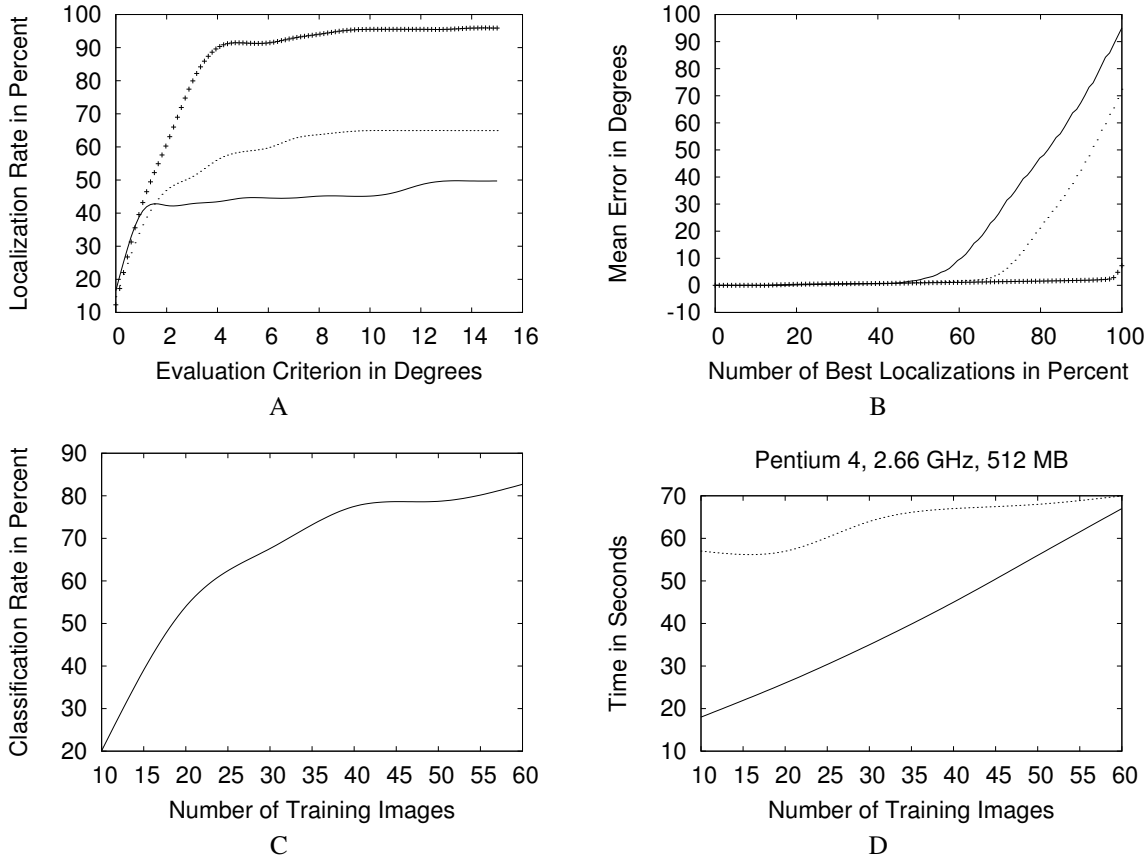

Figure 4. A - localization rate depending on the evaluation criterion for the external rotation (— 20 training images for each class, ··· 40 training images for each class, +++ 60 training images for each class); B - mean localization error depending on the number of best localizations (— 20 training images for each class, ··· 40 training images for each class, +++ 60 training images for each class); C - classification rate depending on the number of training images; D - execution time depending on the number of training images (— creation of one object model with the preprocessing step, ··· recognition in 500 test images).

We created 6 object models for each object class using 10, 20, 30, 40, 50, and 60 training images. In this way we can show how the recognition rate depends on the number of training images.

In the recognition phase images with one, two, three or four objects on heterogeneous background were used. Three example test images can be seen in the figure 3J, 3K, 3L. We took 36 test images of each object class separately from one another. Then we created for each of the 36 positions $\binom{9}{1} = 9$ scenes with one object, $\binom{9}{2} = 36$ scenes with two objects, $\binom{9}{3} = 84$ scenes with three objects, and $\binom{9}{4} = 126$ scenes with four objects. We used altogether 9180 multi-object scenes with heterogeneous background. The images used in the recognition phase and the training images were disjunctive. The positions of objects in the test images were not the same as in the training phase. In our experiments the illumination conditions were not constant, which proves the illumination independency of our system.

The robustness of the system was evaluated in many ways. First it was analyzed in how many cases the number of objects in the images was correctly estimated depending on the abort criterion (table 1). We can generally say that the smaller the criterion is, the more frequently too many objects in a scene are found. The higher the criterion is, the more frequently too few objects in a scene are found. Then we computed the classification rates only in the images with correctly estimated number of objects. In figure 4C the classification rate depending on the number of training images can be seen. In 82,7% of the cases the classification was correct using 60 training images for each object class. The localization rate is presented as a function of the evaluation criterion for the localization (maximum allowed deviation of the angle). Figure 4A shows that the more training im-

S_p	-2 obj.	-1 obj.	OK	+1 obj.	+2 obj.
0.40	0%	0%	82%	14%	4%
0.45	0%	0%	85%	12%	3%
0.50	0%	0%	91%	7%	2%
0.55	0%	0%	95%	3%	2%
0.60	0%	3%	97%	0%	0%
0.65	1%	5%	94%	0%	0%
0.70	3%	13%	84%	0%	0%

Table 1. Estimation of number of objects in recognition scenes depending on the abort criterion S_p."+1 obj." means that one object too much in the scene was found.

ages are used, the better the localization works. In the figure 4B the mean localization error depending on the number of best localizations is depicted. We evaluated also the execution time of the training and recognition phase (figure 4D). The preprocessing of 60 training images and creation of one object model takes 67s on a pentium 4, 2.66 GHz. The recognition in 500 test images takes 70s on the same machine.

5. Conclusions

In this paper, we presented an approach for the statistical, appearance-based object recognition of 3-D objects in 2-D gray level images with multi-objects. At the beginning, we described the whole training phase with all its components. The most important innovation compared to [9] is the serial search algorithm for statistical object recognition in 3-D multi-object scenes. The algorithm is able to classify, localize and estimate the number of objects in the scene.

In the future we will extend this approach and combine it with the context modeling of object correspondences in multi-object images (e.g. with Bayesian networks). We will also evaluate our algorithm with partly occluded objects and accelerate the serial search algorithm in some places.

References

[1] R. Brunelli and T. Poggio. Template matching: Matched spatial filters and beyond. 30(5):751–768, Mai 1997.

[2] C. K. Chui. *An Introduction to Wavelets*. ACP, San Diego, 1992.

[3] V. C. de Verdiere and J. L. Crowley. Visual recognition using local appearance. In *Fifth European Conference on Computer Vision (ECCV)*, pages 640–654, Freiburg, Juni 1998.

[4] T. Deselaers, D. Keysers, and H. Ney. Local representation for multi-object recognition. In B. Michaelis and G. Krell, editors, *Pattern Recognition, 25rd DAGM Symposium*, pages 305–312, Magdeburg, Germany, September 2003. Springer-Verlag, Berlin, Heidelberg, New York.

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

[6] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, Juli 1989.

[7] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(2):5–24, 1995.

[8] J. Pösl. *Erscheinungsbasierte statistische Objekterkennung*. Shaker Verlag, Aachen, 1999.

[9] M. Reinhold, D. Paulus, and H. Niemann. Appearance-Based Statistical Object Recognition by Heterogenous Background and Occlusions. In B. Radig and S. Florczyk, editors, *Pattern Recognition, 23rd DAGM Symposium*, pages 254–261, München, September 2001. Springer-Verlag, Berlin, Heidelberg, New York. Lecture Notes in Computer Science 2191.

[10] M. Reinhold, D. Paulus, and H. Niemann. Improved Appearance-Based 3-D Object Recognition Using Wavelet Features. In T. Ertl, B. Girod, G. Greiner, H. Niemann, and H.-P. Seidel, editors, *Vision, Modeling, and Visualization 2001*, pages 473–480, Stuttgart, November 2001. AKA/IOS Press, Berlin, Amsterdam.

[11] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.