

M. Grzegorzek and E. Izquierdo. Statistical 3d object classification and localization with context modeling. In M. Domanski, R. Stasinski, and M. Bartkowiak, editors, *15th European Signal Processing Conference*, pages 1585–1589, Poznan, Poland, September 2007. PTETiS, Poznan.

STATISTICAL 3D OBJECT CLASSIFICATION AND LOCALIZATION WITH CONTEXT MODELING

Marcin Grzegorzek and Ebroul Izquierdo

Multimedia & Vision Research Group
Queen Mary, University of London
Mile End Road, E1 4NS London, UK
marcin.grzegorzek@elec.qmul.ac.uk

ABSTRACT

This contribution presents a probabilistic approach for automatic classification and localization of 3D objects in 2D multi-object images taken from a real world environment. In the training phase, statistical object models and statistical context models are learned separately. For the object modeling, the recognition system extracts local feature vectors from training images using the wavelet transformation and models them statistically by density functions. Since in contextual environments a-priori probabilities for occurrence of different objects cannot be assumed to be equal, statistical context modeling is introduced in this work. The a-priori occurrence probabilities are learned in the training phase and stored in so-called context models. In the recognition phase, the system determines the unknown number of objects in a multi-object scene first. Then, the object classification and localization are performed. Recognition results for experiments made on a real dataset with 3240 test images compare the performance of the system with and without consideration of the context modeling.

1. INTRODUCTION

One of the most fundamental problems of computer vision is the recognition of objects in digital images [9]. The term object recognition comprehends both, classification and localization of objects. The task of object classification is to determine the classes of objects occurring in the image \mathbf{f} from a set of predefined object classes $\Omega = \{\Omega_1, \Omega_2, \dots, \Omega_K, \dots, \Omega_{N_\Omega}\}$. Generally, the number of objects in a scene is unknown. Therefore, it is necessary to find out the number of objects in the image first. In the case of object localization, the recognition system estimates the poses of objects in the image, whereas the object classes are assumed to be a-priori known. The object poses are defined relatively to each other with a 3D translation vector $\mathbf{t} = (t_x, t_y, t_z)^T$ and a 3D rotation vector $\boldsymbol{\phi} = (\phi_x, \phi_y, \phi_z)^T$ in a coordinate system with an origin placed in the image center.

There are two main approaches for object recognition, namely shape-based and appearance-based methods. The shape-based algorithms perform a segmentation and use geometric features like lines or corners for object representation [2, 5]. Unfortunately, these methods suffer often from segmentation errors. Therefore, many authors, e.g., [8, 12], prefer a second method, the appearance-based object recognition. Here, texture is taken into consideration

for object description. The object features are computed directly from the pixel values without a previous segmentation step. Most fundamental approaches for appearance-based object classification and localization are template matching [1, 11], eigenspace approach [7] and Support Vector Machines [3, 14].

Many approaches for automatic object recognition do not take any context information of a scene into account, e.g., [10, 12]. These algorithms easily assume that the a-priori occurrence probabilities for all object classes Ω_K considered in a particular recognition task are equal. However, having additional knowledge about the environment, in which a scene was taken, the occurrence of some objects might be more likely than the occurrence of the others [4]. Considering this additional knowledge in the learning phase is called context modeling. Figure 1 shows three example contexts, namely the office context, the kitchen context, and the nursery context. In the office context, objects like punchers, staplers, or pens can be found more likely than, e.g., plates, knives, or forks, which are rather expected in the kitchen. Therefore, it is useful to model the context dependencies between objects in the training phase.

In the present work, statistical context modeling for multi-object scenes is introduced. Here, the a-priori occurrence probabilities are not assumed to be equal for all objects. They are learned in the training phase for each context separately using an additional and very large training dataset. Moreover, the system described in this contribution extracts local feature vectors directly from pixel intensities (appearance-based approach) using the wavelet multiresolution analysis [6] and models them by density functions (statistical recognition [15]).

This paper is structured as follows. Section 2 presents the training of statistical object models and statistical context models. In Section 3, the recognition phase of the system is discussed. The system performance with and without consideration of the context modeling is compared for experiments made on a real dataset with 3240 test images in Section 4. Section 5 closes this contribution with some final remarks and a conclusion.

2. STATISTICAL MODELING

Before objects can be classified and localized in the recognition phase (Section 3), object models \mathcal{M}_K for all object classes Ω_K considered in a particular recognition task are learned in the training phase (Section 2.1). Moreover, context dependencies between objects are statistically modeled in order to improve recognition rates for multi-object scenes (Section 2.2).

The research activity leading to this work has been supported by the European Commission under the contract FP6-027026-K-SPACE.



Figure 1: Left: office context. Middle: kitchen context. Right: nursery context.

2.1 Object Modeling

The object modeling starts with the collection of training data performed by a special setup with a turntable and camera arm. Under training data for the object modeling, both the images $\mathbf{f}_{\kappa, \rho=1, \dots, N_p}$ of the objects and the object poses $(\phi_{\kappa, \rho}, \mathbf{t}_{\kappa, \rho})$ in these images are understood. Subsequently, the original training images are converted and resized into gray level images of size $2^n \times 2^n$ ($n \in \mathbb{N}$) pixels.

In all these preprocessed training images 2D local feature vectors $\mathbf{c}_{\kappa, m}$ are extracted using the wavelet transformation [6]. Training images are divided into neighborhoods of size $2^{|\hat{s}|} \times 2^{|\hat{s}|}$ (in Figure 2, 4×4 pixels). These neighborhoods are treated as 2D discrete signals b_0 and decomposed to low-pass and high-pass coefficients. The resulting coefficients $b_{\hat{s}}$, $d_{0, \hat{s}}$, $d_{1, \hat{s}}$, and $d_{2, \hat{s}}$ are then used for feature vector computation

$$\mathbf{c}_{\kappa, m}(\mathbf{x}_m) = \begin{pmatrix} \ln(2^{|\hat{s}|} b_{\hat{s}}) \\ \ln[2^{|\hat{s}|} (|d_{0, \hat{s}}| + |d_{1, \hat{s}}| + |d_{2, \hat{s}}|)] \end{pmatrix}. \quad (1)$$

As one can imagine, some feature vectors in each training image describe the object, others belong to the background. In real world environment, it cannot be assumed that the background in the recognition phase is a-priori known. Therefore, only feature vectors describing the object should be considered for the statistical object modeling. Since the object takes usually only a part of the image, a tightly enclosing object area (bounding region) O_κ for each object class Ω_κ is defined. This object area can be regarded as a function $O_\kappa(\phi, \mathbf{t})$ defined on a continuous pose parameter domain (ϕ, \mathbf{t}) for all object classes Ω_κ . For each object class Ω_κ and pose parameters (ϕ, \mathbf{t}) , it determines the set C_{O_κ} of feature vectors $\mathbf{c}_{\kappa, m}$ describing the object. The remaining feature vectors $\mathbf{c}_{\kappa, m} \notin C_{O_\kappa}$ are called background features.

The object feature vectors $\mathbf{c}_{\kappa, m} \in C_{O_\kappa}$ are modeled by normal density functions $p(\mathbf{c}_{\kappa, m} | \boldsymbol{\mu}_{\kappa, m}, \boldsymbol{\sigma}_{\kappa, m}, \phi, \mathbf{t})$, whereas the corresponding mean value vectors $\boldsymbol{\mu}_{\kappa, m}$ are represented as functions $\boldsymbol{\mu}_{\kappa, m}(\phi, \mathbf{t})$; while standard deviation vectors $\boldsymbol{\sigma}_{\kappa, m}$ are modeled with constant components.

Finally, statistical object models $\mathcal{M}_\kappa(\phi, \mathbf{t})$ for all object classes Ω_κ are created. These object models are considered as continuous functions of the transformation parameters (ϕ, \mathbf{t}) .

2.2 Context Modeling

In contextual environments, a-priori occurrence probabilities cannot be assumed to be equal for all object classes (see Figure 1). They have to be learned in the training phase. First,

the set Υ of contexts $\Upsilon_{l=1, \dots, N_\Upsilon}$ considered in a particular object recognition task is introduced

$$\Upsilon = \{\Upsilon_1, \Upsilon_2, \dots, \Upsilon_l, \dots, \Upsilon_{N_\Upsilon}\} \quad (2)$$

It is assumed that the number N_Υ and the kinds (kitchen, bathroom, etc.) of the contexts are known. Moreover, the set of object classes $\Omega = \{\Omega_1, \Omega_2, \dots, \Omega_\kappa, \dots, \Omega_{N_\Omega}\}$ is also known for the learning of the context dependencies. The training of the context dependencies between objects starts with the image acquisition. First, N_l images from random viewpoints are taken with a hand-held camera for each context Υ_l . Second, it is manually determined, which of the objects $\Omega_{\kappa=1, \dots, N_\Omega}$ and how often occur in the images, whereas with $N_{l, \kappa}$ the number is denoted, how often the object Ω_κ occurs in the context Υ_l . Generally, the sum of $N_{l, \kappa}$ for all object classes $\Omega_{\kappa=1, \dots, N_\Omega}$ is not equal to N_l . Therefore, for all contexts $\Upsilon_{l=1, \dots, N_\Upsilon}$ a normalization factor η_l is introduced so that

$$\eta_l(N_{l, 1} + N_{l, 2} + \dots + N_{l, \kappa} + \dots + N_{l, N_\Omega}) = N_l. \quad (3)$$

Using this normalization factor η_l and the number $N_{l, \kappa}$, the a-priori occurrence probability for the object Ω_κ in the context Υ_l is learned as

$$p_l(\Omega_\kappa) = \eta_l N_{l, \kappa} \quad (4)$$

These a-priori probabilities stored in statistical context models \mathcal{M}_l are used in the recognition phase for multi-object scenes with context dependencies (Section 3.3).

3. OBJECT RECOGNITION

Once the object modeling (Section 2.1) and the context modeling (Section 2.2) are finished, the system is able to classify and localize objects in images taken from a real world contextual environment. First, a test image \mathbf{f} is taken, pre-processed, and local feature vectors \mathbf{c}_m in it are computed in the same way as in the training phase (Section 2.1). Second, one of the recognition algorithms integrated into the system is started. The classification and localization algorithm for single-object scenes is described in Section 3.1. Its extension to multi-object scenes without context modeling follows in Section 3.2. In Section 3.3, the context dependencies in multi-object scenes are additionally taken into consideration.

3.1 Single-Object Scenes

The task of the classification and localization algorithm for single-object scenes is to find the class $\Omega_{\hat{\kappa}}$, (or just its index $\hat{\kappa}$) and the pose $(\hat{\phi}, \hat{\mathbf{t}})$ of the object, which occurs in the test

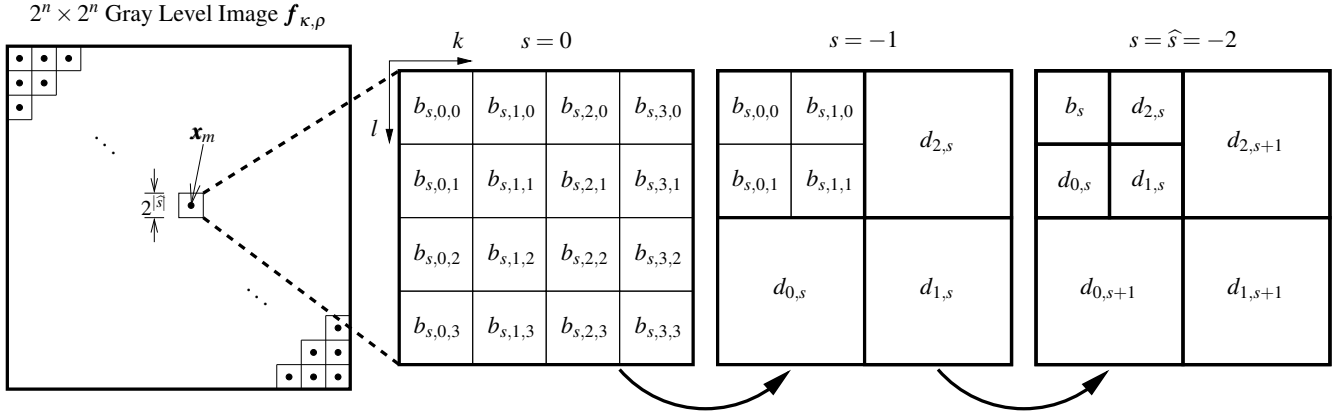


Figure 2: 2D signal decomposition with the wavelet transformation for a local neighborhood of size 4×4 pixels. The final coefficients result from gray values $b_{0,k,l}$ and have the following meaning: b_{-2} : low-pass horizontal and low-pass vertical, $d_{0,-2}$: low-pass horizontal and high-pass vertical, $d_{1,-2}$: high-pass horizontal and high-pass vertical, $d_{2,-2}$: high-pass horizontal and low-pass vertical.

image f . In order to do so, the object density values for all objects Ω_κ and many pose hypotheses (ϕ_h, t_h) have to be compared to each other. Assuming that the object feature vectors $\mathbf{c}_m \in C_{O_\kappa}$ are statistically independent on each other, the object density value for the given test image f , object class hypothesis Ω_κ , and object pose hypothesis (ϕ_h, t_h) is computed with

$$p_{\kappa,h} = \prod_{\mathbf{c}_m \in C_{O_\kappa}} p(\mathbf{c}_m | \mu_{\kappa,m}, \sigma_{\kappa,m}, \phi_h, t_h) \quad (5)$$

All data required for computation of the density value $p_{\kappa,h}$ with (5) is stored in the statistical object model $\mathcal{M}_\kappa(\phi_h, t_h)$. These object density values are then maximized with the maximum likelihood (ML) estimation [15]

$$(\hat{\kappa}, \hat{h}) = \underset{(\kappa,h)}{\operatorname{argmax}} p_{\kappa,h} \quad (6)$$

Having the index $\hat{\kappa}$ of the resulting class and the index \hat{h} of the resulting pose hypothesis, the classification and localization problem for the single-object scene f is solved.

3.2 Multi-Object Scenes without Context

The recognition algorithm for multi-object scenes without consideration of the context dependencies assumes the uniform distribution (??) of the a-priori occurrence probabilities for all object classes. In the recognition task for multi-object scenes, not only the classes of objects and their poses have to be determined. Since the number of objects in a scene is a-priori unknown, it also must be estimated. The initial point for this algorithm is the recognition approach for single-object scenes presented in Section 3.1. First, the ML algorithm estimates the optimal pose parameters $(\hat{\phi}_\kappa, \hat{t}_\kappa)$ for all object classes Ω_κ considered in the recognition task by maximizing the object density value according to (6). This

can be expressed with the following maximization terms

$$\begin{aligned} (\hat{h}_1) &= \underset{(h)}{\operatorname{argmax}} p_{1,h} \\ &\dots \\ (\hat{h}_\kappa) &= \underset{(h)}{\operatorname{argmax}} p_{\kappa,h} \quad (7) \\ &\dots \\ (\hat{h}_{N_\Omega}) &= \underset{(h)}{\operatorname{argmax}} p_{N_\Omega,h} \end{aligned}$$

The object density values for the optimal pose hypotheses can be written in short forms as follows

$$\begin{aligned} \hat{Q}_1 &= p_{1,\hat{h}} \\ &\dots \\ \hat{Q}_\kappa &= p_{\kappa,\hat{h}} \quad (8) \\ &\dots \\ \hat{Q}_{N_\Omega} &= p_{\kappa,\hat{N}_\Omega} \end{aligned}$$

These object densities $\hat{Q}_{\kappa=1,\dots,N_\Omega}$ are now sorted from the highest to the lowest value, i. e., in a non-increasing way

$$\underbrace{\hat{Q}_{\kappa_1} \geq \hat{Q}_{\kappa_2}}_{d_1} \geq \dots \geq \underbrace{\hat{Q}_{\kappa_i} \geq \hat{Q}_{\kappa_{i+1}}}_{d_i} \geq \dots \geq \hat{Q}_{\kappa_I} \quad (9)$$

where $I = N_\Omega$ and d_i is a difference between neighboring elements

$$d_i = d(\hat{Q}_{\kappa_i}, \hat{Q}_{\kappa_{i+1}}) = \hat{Q}_{\kappa_i} - \hat{Q}_{\kappa_{i+1}} \quad (10)$$

Finally, the index \hat{i} of the highest distance $d_{\hat{i}} (\forall i \neq \hat{i} : d_i \leq d_{\hat{i}})$ can be easily estimated with the following formula

$$\hat{i} = \underset{i}{\operatorname{argmax}} d_i \quad (11)$$

and is interpreted as the number of objects occurring in the multi-object scene f . Hence, the final recognition result in the multi-object scene f are the following object classes and

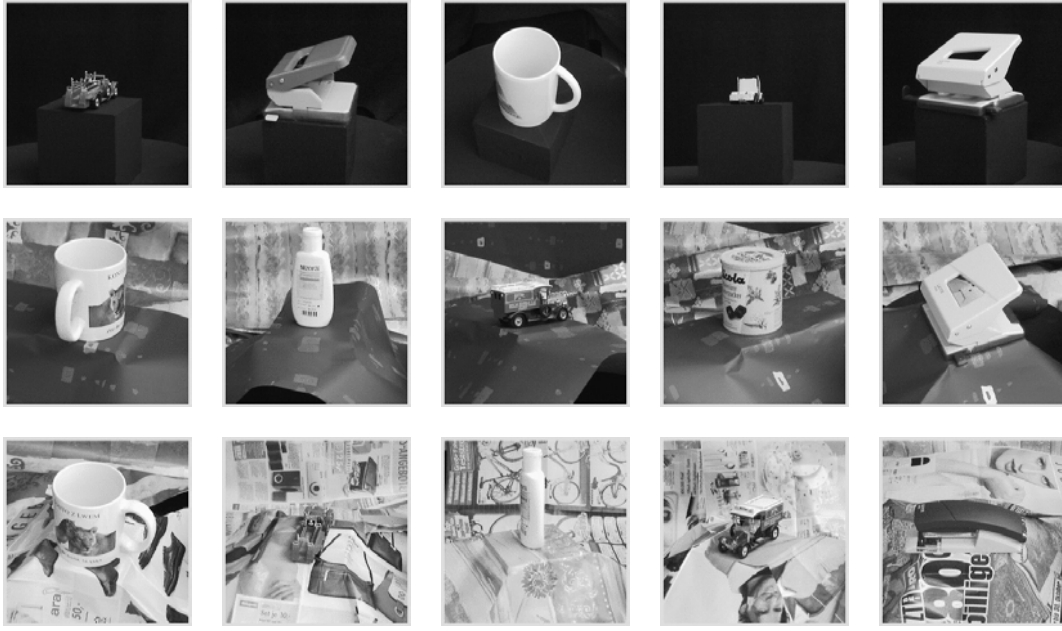


Figure 3: Example single-object test images from the 3D-REAL-ENV database [13]. First row: test images with homogeneous background. Second row: test images with less heterogeneous background. Third row: test images with more heterogeneous background.

poses

$$\begin{array}{ll}
 \text{first object} & (\kappa_1, \hat{\phi}_{\kappa_1}, \hat{t}_{\kappa_1}) \\
 \text{second object} & (\kappa_2, \hat{\phi}_{\kappa_2}, \hat{t}_{\kappa_2}) \\
 & \vdots \\
 \text{last object} & (\kappa_{\hat{i}}, \hat{\phi}_{\kappa_{\hat{i}}}, \hat{t}_{\kappa_{\hat{i}}})
 \end{array} \quad (12)$$

In order to evaluate the recognition algorithm for multi-object scenes, not only the object classification result Ω_{κ_i} and the object localization result $(\hat{\phi}_{\kappa_i}, \hat{t}_{\kappa_i})$ are verified. The number \hat{i} of objects found in the scene \mathbf{f} is also checked (Section 4).

3.3 Multi-Object Scenes with Context

The recognition algorithm for multi-object scenes with context dependencies uses the context models \mathcal{M}_i learned as shown in Section 2.2. Searching for the first object Ω_{κ_1} in the multi-object scene \mathbf{f} , the algorithm does not use any context information and, similarly to previous sections, it assumes equal a-priori probabilities (??) for all object classes $\Omega_{\kappa=1, \dots, N_{\Omega}}$ considered in the recognition task. The class κ_1 and the pose $(\hat{\phi}_1, \hat{t}_1)$ of the first object in the image \mathbf{f} is determined in the same way as in Section 3.1. Subsequently, the context $\Upsilon_{\hat{i}}$ for the scene \mathbf{f} (or just the context number \hat{i}) is determined using the statistical context models $\mathcal{M}_{i=1, \dots, N_{\Upsilon}}$. Each context model \mathcal{M}_i contains the trained a-priori probabilities $p_i(\Omega_{\kappa})$ for all object classes $\Omega_{\kappa=1, \dots, N_{\Omega}}$. Therefore, using the context models $\mathcal{M}_{i=1, \dots, N_{\Upsilon}}$ it is possible to determine the a-priori density $p_{i=1, \dots, N_{\Upsilon}}(\Omega_{\kappa_1})$ of the first object class Ω_{κ_1} for all contexts $\Upsilon_{i=1, \dots, N_{\Upsilon}}$. The highest value of these densities $p_{i=1, \dots, N_{\Upsilon}}(\Omega_{\kappa_1})$ decides about the context $\Upsilon_{\hat{i}}$

of the multi-object scene \mathbf{f}

$$\hat{i} = \underset{i}{\operatorname{argmax}} p_i(\Omega_{\kappa_1}) \quad (13)$$

Looking for further objects $(\Omega_{\kappa_2}, \Omega_{\kappa_3}, \dots, \Omega_{\kappa_{\hat{i}}})$ in the image \mathbf{f} the statistical context model $\mathcal{M}_{\hat{i}}$ learned for the context $\Upsilon_{\hat{i}}$ is used and the following a-priori probabilities for object occurrence

$$p_{\hat{i}}(\Omega_1) \neq \dots \neq p_{\hat{i}}(\Omega_{\kappa}) \neq \dots \neq p_{\hat{i}}(\Omega_{N_{\Omega}}) \quad (14)$$

are taken into consideration. Further procedure for object classification and localization is almost identical to object recognition for multi-object scenes without context (Section 3.2). First, the maximum likelihood estimation is applied according to (7). Second, the object density values for the optimal pose hypotheses are written in short forms

$$\begin{aligned}
 \hat{Q}_1 &= p_{\hat{i}}(\Omega_1) p_{1, \hat{h}} \\
 &\dots \\
 \hat{Q}_{\kappa} &= p_{\hat{i}}(\Omega_{\kappa}) p_{\kappa, \hat{h}} \quad , \\
 &\dots \\
 \hat{Q}_{N_{\Omega}} &= p_{\hat{i}}(\Omega_{N_{\Omega}}) p_{\kappa, \hat{N}_{\Omega}}
 \end{aligned} \quad (15)$$

whereas here they are weighted by the a-priori probabilities stored in the statistical context model $\mathcal{M}_{\hat{i}}$. Subsequently, the weighted object densities $\hat{Q}_{\kappa=1, \dots, N_{\Omega}}$ are sorted in a non-increasing way (10). Finally, the index \hat{i} of the highest distance $d_{\hat{i}}$ is estimated with (11) and the classification and localization result can be presented with (12).

4. EXPERIMENTS AND RESULTS

In the testing phase of the recognition algorithms for multi-object scenes introduced in Sections 3.2 and 3.3, altogether

3D-REAL-ENV Image Database	Without Context Modeling			With Context Modeling		
	HomBack	LessHetBack	MoreHetBack	HomBack	LessHetBack	MoreHetBack
ObjNumDet	100%	83.9%	43.2%	99.9%	88.2%	59.2%
Classification	100%	91.9%	62.9%	100%	97.0%	87.5%
Localization	99.7%	81.7%	58.1%	99.7%	81.7%	58.1%

Table 1: Quantitative comparison of the system performance with and without context modeling. ObjNumDet - object number determination rate. Classification - classification rate. Localization - localization rate.

3240 gray level multi-object scenes sized 512×512 pixels were used. They were generated based on the single-object test images from the 3D-REAL-ENV image database [13], which consists of 10 objects (examples in Figure 3). The test images can be divided into three types, i.e., there are 1080 multi-object scenes with homogeneous, 1080 multi-object scenes with less heterogeneous, and 1080 multi-object scenes with more heterogeneous background. Additionally, the 3D-REAL-ENV objects (see Figure 3) were assigned into three different contexts, namely the kitchen Υ_1 , the nursery Υ_2 , and the office Υ_3 . For each image type ($T_{\text{type}} \in \{\text{hom}, \text{less}, \text{more}\}$) and each context ($\Upsilon = \{\text{kitchen}, \text{nursery}, \text{office}\}$) 120 one-object, 120 two-object, and 120 three-object scenes were created, whereas the viewpoints were chosen randomly from all 288 3D-REAL-ENV test views [13] and are different for all combinations of the test scenes. For example, two-object test scenes with homogeneous background ($T_{\text{type}} = \text{hom}$) from the kitchen context ($\Upsilon_1 = \text{kitchen}$) represent, in general, different viewpoints as the three-object test scenes with less heterogeneous background ($T_{\text{type}} = \text{less}$) from the office context ($\Upsilon_3 = \text{office}$). The quantitative comparison of the system performance with and without context modeling is presented in Table 1. Since the localization is performed for a-priori known object classes, the context modeling does not influence its rate.

5. CONCLUSION

In this paper, a statistical recognition system for multi-object scenes with context dependencies was presented. Since in contextual environments a-priori probabilities for occurrence of different objects cannot be assumed to be equal, statistical context modeling was introduced in this work (Section 2.2). Recognition results achieved for experiments presented in Section 4 prove a very high performance of the system in a real world environment. Due to the main contribution of this article, statistical context modeling, the classification rate increased from 62.9% to 87.5%.

In the future, the appearance-based approach described in this work will be combined with the shape-based method for object recognition. There are objects with the same shape, which are distinguishable only by texture, but one can imagine also objects with the same texture features, which are easy to distinguish by shape.

REFERENCES

- [1] R. Brunelli and T. Poggio. Template matching: Matched spatial filters and beyond. *Pattern Recognition*, 30(5):751–768, May 1997.
- [2] H. Chen, I. Shimshoni, and P. Meer. Model based object recognition by robust information fusion. In *17th International Conference on Pattern Recognition*, Cambridge, UK, August 2004.
- [3] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.
- [4] M. Grzegorzec. *Appearance-Based Statistical Object Recognition Including Color and Context Modeling*. Logos Verlag, Berlin, Germany, 2007.
- [5] L. J. Latecki and R. Lakaemper. Application of planar shape comparison to object retrieval in image databases. *Pattern Recognition*, 35(1):15–19, January 2002.
- [6] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.
- [7] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *PAMI*, 19(7):696–710, Juli 1997.
- [8] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, January 1995.
- [9] H. Niemann. *Pattern Analysis and Understanding*. Springer-Verlag, Berlin, Heidelberg, Germany, 1990.
- [10] J. Pösl. *Erscheinungsbasierte, statistische Objekterkennung*. Shaker Verlag, Aachen, Germany, 1999.
- [11] W. K. Pratt. *Digital Image Processing*. John Wiley & Sons Ltd, New York, USA, 2001.
- [12] M. Reinhold. *Robuste, probabilistische, erscheinungsbasierte Objekterkennung*. Logos Verlag, Berlin, Germany, 2004.
- [13] M. Reinhold, M. Grzegorzec, J. Denzler, and H. Niemann. Appearance-based recognition of 3-d objects by cluttered background and occlusions. *Pattern Recognition*, 38(5):739–753, May 2005.
- [14] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, USA, 1995.
- [15] A. R. Webb. *Statistical Pattern Recognition*. John Wiley & Sons Ltd, Chichester, UK, 2002.