Statistical 3–D Object Localization Without Segmentation Using Wavelet Analysis

Josef Pösl and **Heinrich Niemann**

Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg
Martensstr. 3, D–91058 Erlangen, Germany
email: {poesl,niemann}@informatik.uni-erlangen.de

Abstract. This paper presents a new approach for statistical object localization. The localization scheme is directely based on local features, which are extracted for all image positions, in contrast to segmentation in classical schemes. Hierarchical Gabor filters are used to extract local features. With these features statistical object models are built for the different scale levels of the Gabor filters. The localization is then performed by a maximum likelihood estimation on the different scales successively. Results for the localization of real images of 2–D and 3–D objects are shown.

1 Introduction and Motivation

When analysing a 2–D gray–level image with multiple 3–D objects two major problems have to be solved: The pose estimation and classification of each object in the scene. In this paper we focus on the localization of one individual object. All other objects are considered as belonging to the background. We define a statistical density function for pose estimation.

Recognition results in speech understanding were strongly enhanced by the idea of incorporating statistical methods in the recognition task. Despite the obvious success in this area a statistical framework for image object recognition has not been investigated widely up to now. Nevertheless, recent results prove this approach as promising [3].

Most publications in the field of statistical object modeling use geometric information of segmentation results as random variables. Lines or vertices, for example, can be used for the construction of statistical models. There are two major disadvantages of using solely segmentation results. When restricting the recognition process to this level of abstraction a lot of information contained in an image is lost. Another disadvantage are the errors made by the segmentation.

One way to cope with the problem is to avoid segmentation. Instead of that the gray–level information of an image can be used. Correlation is the simplest method of doing this. Another approach focuses in singular value decompositions of vector spaces composed of the gray–level data of several images (appearance based modeling) [1, 8]. Maximization of the mutual information between an object model and an object in a scene is a further possibility [10]. A similar technique is described in [5]. [6] describes a method based on mixture densities of

the gray level values of object images. With a focus on the distributions of image pixel values rather than object location values and without an hierarchical solution this approach tends to be very complex.

The cited papers either use only probabilistically chosen model points for matching [10] or use pose restrictions [5] to reduce the complexity of the estimation. In this paper a new approach for the localization of 3–D objects in single gray–level images is presented. The pose of the object is not restricted and the complete image data is considered after hierarchical filtering. Local features are modeled statistically. We demonstrate a new way of formulating a statistical model with a functional basis decomposition of probability density parameters.

2 System overview

The aim of the presented system is the pose estimation of a rigid 3–D object in a single 2–D gray–level image. The parameter space is six-dimensional for this task. Let $\boldsymbol{R}_x$,$\boldsymbol{R}_y$ and $\boldsymbol{R}_z$ denote the 3D–rotation matrices with rotation angle ϕ_x, ϕ_y and ϕ_z round the x–, y– and z–axis respectively. The 3D–transformation consists of the rotation $\boldsymbol{R} = \boldsymbol{R}_z\boldsymbol{R}_y\boldsymbol{R}_x \in \mathbb{R}^{3\times 3}$ and the translation $\boldsymbol{t} = (t_x, t_y, t_z)^{\mathrm{T}} \in \mathbb{R}^3$. The parameter space can be split into a rotation $\boldsymbol{R}_z$ with angle $\boldsymbol{\phi}_{int} = (\phi_z)$ and a translation $\boldsymbol{t}_{int} = (t_x, t_y, 0)^{\mathrm{T}}$ inside the image plane and orthogonal components $\boldsymbol{R}_y\boldsymbol{R}_x$ ($\boldsymbol{\phi}_{ext} = (\phi_y, \phi_x)$) and $\boldsymbol{t}_{ext} = (0, 0, t_z)^{\mathrm{T}}$ for the transformations outside. For this work it is assumed that the object does not vary in scale: $\boldsymbol{t} = \boldsymbol{t}_{int}$.

In a first step of the localization process a multiresolution analysis of the image is used to derive feature values on different scales $s \in \mathbb{Z}$ and resolutions (sampling rates) $\Delta x_s = \Delta y_s = r_s \in \mathbb{R}^+$ at the locations of rectangular sampling grids ($r_{s,q+1} < r_{s,q}$). The image $f(x, y)$ is transformed to signals $\boldsymbol{h}_s = (h_{s,0}, \ldots, h_{s,N-1})^{\mathrm{T}}$ by local transformations $\mathcal{T}_{s,n}$ for scale s: $h_{s,n}(x, y) = \mathcal{T}_{s,n}\{f\}(x, y)$. Feature vectors $\boldsymbol{c}_{s,k,l} = (c_{s,k,l,0}, \ldots, c_{s,k,l,N-1})^{\mathrm{T}}$ at discrete locations are obtained by subsampling: $c_{s,k,l,n} = \mathcal{T}_{s,n}\{f\}(kr_s, lr_s)$. Possible definitions of this transformation are described in section 3.

We define a statistical measure for the probability of those features under the assumption of an object transformation. The complexity of the pose estimation is high if all features on the different scale levels are combined into one measure function. Therefore, a hierarchical solution is used. Measures are defined for each scale. The analysis starts on a low scale and a rough resolution. The resolution is then decreased step by step. The transformation estimation becomes more exact with each step. Let $\tilde{\boldsymbol{c}}_s$ be the vector of the concatenated feature vectors detected in an image on scale s, $\boldsymbol{B}_s$ the model parameters of an object class and $\boldsymbol{R}, \boldsymbol{t}$ be parameters for rotation and translation. The model parameters $\boldsymbol{B}_s$ consist of geometric information like probability density locations and other density parameters. The density $p(\tilde{\boldsymbol{c}}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t})$ is then used for localization. The maximum likelihood estimation results in $(\widehat{\boldsymbol{R}}_s, \widehat{\boldsymbol{t}}_s) = \mathrm{argmax}_{(\boldsymbol{R},\boldsymbol{t})}\, p(\tilde{\boldsymbol{c}}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t})$.

Given a descending sequence $(s_q)_{q=0,\ldots,N_s-1}$ of scales, the analysis begins with the roughest scale s_0. The parameter space is searched completely at this

level. The best H transformation hypotheses on scale level s_q are then used to perform a search with restricted parameter range on scale level s_{q+1}. The optimum search on all levels consists of a combination of a grid search for the given parameter range and a successive local search. The grid search on level s_0 is a global search of the full parameter range while the grid search on each other level evaluates the optimum only in the neighbourhood of the previous local optimum. The grid resolution thereby decreases with the scale levels (Fig. 1).

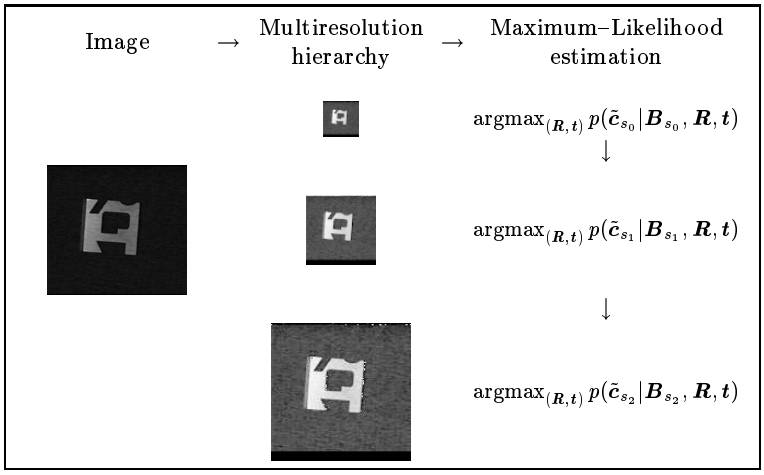

Fig. 1. System overview: Probability density maximization on multiresolution hierarchy of images.

3 Features

Features used for object recognition and localization must have at least two properties. First, they should be robust with respect to noise and different lighting conditions in a scene. Second, they should be invariant to the transformations of the object in the image plane. The feature values at certain object locations should not change if the object is translated or rotated in the image plane.

We define local features derived from Gabor filters which fulfil the requirements of robustness and invariance. By suppressing high frequencies at different scales Gabor filters are a means to compensate for noise. Gabor functions are Gaussians modulated by complex sinusoids. The 2–D Gabor function can be written as:

$$g(\boldsymbol{x}, \boldsymbol{\omega}) = \exp\left(-\left[\frac{x^2}{2{\sigma_x}^2} + \frac{y^2}{2{\sigma_y}^2}\right] + 2\pi i\left[\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}\right]\right).$$

$\boldsymbol{\sigma} = (\sigma_x, \sigma_y)^{\mathrm{T}}$ defines the width of the Gaussian in the spatial domain and $\boldsymbol{\omega}$ is the frequency of the complex sinusoid. These functions achieve the minimum possible joint resolution in space and frequency [2]. They furthermore form a complete but nonorthogonal basis for the set of two-dimensional functions. In order to derive suitable features from the filter results, Gabor wavelets

are defined [9]. They allow the analysis of the signal at different scale levels and vary spatial and frequency windows accordingly. The *basic wavelet* of the Gabor wavelet transform with circular spatial windows ($\sigma_x = \sigma_y$) is defined as:

$$g(\boldsymbol{x}, \theta) = \exp\left(-\boldsymbol{x}^2 + p\pi i x'\right),$$

where θ is the orientation and $\boldsymbol{x}' = (x', y')^{\mathrm{T}} = \boldsymbol{R}_\theta \boldsymbol{x}$. $\boldsymbol{R}_\theta$ denotes the 2–D rotation matrix with angle θ. The constant p which specifies the ratio of spatial extent and period of the functions is chosen as $p = 1$ according to [7]. With these definitions the wavelet transform of a 2–D function f on the scale $s \in \mathbb{Z}$ is:

$$w_s(\boldsymbol{x}, \theta) = \int f(\boldsymbol{x}_0) \bar{g}\left(d^{-s}(\boldsymbol{x} - \boldsymbol{x}_0), \theta\right) d\boldsymbol{x}_0,$$

with $d \in \mathbb{R}$, $d > 1$, $s \in \mathbb{Z}$ and $\theta \in \{\theta_l = \frac{l\pi}{N_l}\}_{l=0,\ldots,N_l-1}$. $\bar{g}$ is the conjugate of g. A feature vector $\boldsymbol{c}_s = (c_{s,0}, \ldots, c_{s,N-1})^{\mathrm{T}} = \boldsymbol{c}_s(\boldsymbol{x})$ can now be defined for the locations $\boldsymbol{x}$ on different scales s:

$$c_{s,n} = \left| FT_{k=1\ldots N-1} \left\{ \left| FT_{l=0..N-1} \left\{ \log |w_s(\boldsymbol{x}, \theta_l)| \right\}_k \right| \right\}_n \right|, N = \left\lfloor \frac{N_l + 1}{2} \right\rfloor,$$

where

$$FT_{k=k_0\ldots k_1}\{f_k\}_l = \sum_{k=k_0\ldots k_1} f_k \exp\left(-\frac{2\pi i k l}{N}\right)$$

is the discrete Fourier transform. It is approximately (asymtotically for $N \to \infty$) rotationally invariant and robust to changes in lighting conditions.

As already stated in Sect. 2 the localization is performed hierarchically on different scale levels. The resolution r_s (sampling distance) of the analysis on scale s is connected to the spatial filter width by a constant λ: $r_s = \lambda d^s$.

4 Statistical model

4.1 Model formulation

This section shows the definition of a probability density function on each of the scale levels of the analysis. To simplify the notation the index s indicating the scale level is omitted.

The model object is covered with a rectangular grid of local feature vectors (see Fig. 2). The grid resolution is the same as the image resolution on the actual scale. Let $A \subset X$ be a small region (e.g. rectangular) which contains the object projection onto the image plane for all possible rotations $\boldsymbol{\phi}_{ext}$ and constant $\boldsymbol{\phi}_{int}$ and $\boldsymbol{t}$ (see Fig. 2). Let $X = \{\boldsymbol{x}_m\}_{m=0,\ldots,M-1}$, $\boldsymbol{x}_m \in \mathbb{R}^2$ denote the grid locations and $\boldsymbol{c}(\boldsymbol{x})$ the observed feature vector at location $\boldsymbol{x} = (x, y)^{\mathrm{T}}$. In this paper we choose the grid locations as the sampling locations of the image transformation: $X = \left\{\boldsymbol{x}_m = (kr_s, lr_s)^{\mathrm{T}}\right\}$. The local features $c_n(\boldsymbol{x}_m)$ are the components of the image feature vector $\tilde{\boldsymbol{c}}$ if the object is not transformed in the image plane: $c_n(\boldsymbol{x}_m) = c_n(kr_s, lr_s) = c_{k,l,n} = c_{m,n}$.

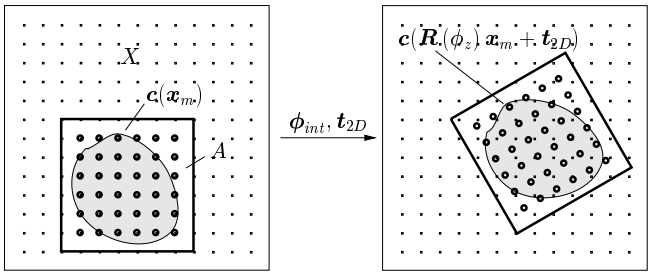

Fig. 2. Object covered with grid for feature extraction.

The local features $c_{m,n}$ are interpreted as random variables. The randomness thereby is, among others, the consequence of noise in the image sampling process and complex changes in environment (e.g. lighting) conditions. Assuming the densities $p(\boldsymbol{c}_m)$ of the local features as stochastically independent leads to:

$$p(\tilde{\boldsymbol{c}}) = \prod_{\boldsymbol{x}_m} p(\boldsymbol{c}(\boldsymbol{x}_m)) = \prod_{\boldsymbol{x}_m \in A} p(\boldsymbol{c}(\boldsymbol{x})|\boldsymbol{x} = \boldsymbol{x}_m) \prod_{\boldsymbol{x}_m \notin A} p(\boldsymbol{c}(\boldsymbol{x})|\boldsymbol{x} = \boldsymbol{x}_m).$$

If a uniform distribution for the features outside the model area A (which belong to background) is assumed the second product in the above equation is constant. So it is sufficient to consider

$$p(\boldsymbol{c}_A) = \prod_{\boldsymbol{x}_m \in A} p(\boldsymbol{c}(\boldsymbol{x})|\boldsymbol{x} = \boldsymbol{x}_m),$$

where $\boldsymbol{c}_A$ is the subvector of $\tilde{\boldsymbol{c}}$ which belongs to A, for pose estimation.

We will first derive the density for the two-dimensional case. We use linear interpolation for reconstruction of $c_n(\boldsymbol{x})$ from the image feature vector $\tilde{\boldsymbol{c}}$.

The grid positions and A are part of the model parameters $\boldsymbol{B}$. If the model is transformed by $(\boldsymbol{\phi}_{int}, \boldsymbol{t})$ in the image plane the density can be written as:

$$p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{\phi}_{int}, \boldsymbol{t}) = \prod_{\boldsymbol{x}_m \in A} p(\boldsymbol{c}(\boldsymbol{x})|\boldsymbol{x} = \boldsymbol{R}(\phi_z)\, \boldsymbol{x}_m + \boldsymbol{t}_{2D}), \quad \boldsymbol{t}_{2D} = (t_x, t_y)^{\mathrm{T}}.$$

The feature vectors are assumed to be normally distributed with independent components. Let $\mathcal{N}(\boldsymbol{c}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ denote the normal densities, where $\boldsymbol{\mu}_m$ is the mean vector and $\boldsymbol{\Sigma}_m$ the covariance matrix of the feature vector $\boldsymbol{c}_m$. In the case of independence, $\boldsymbol{\Sigma}_m$ is a diagonal matrix diag $(\sigma^2_{m,0}, \ldots, \sigma^2_{m,N-1})$. This results in:

$$p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{\phi}_{int}, \boldsymbol{t}_{int}) = \prod_{\boldsymbol{x}_m \in A} \mathcal{N}(\boldsymbol{c}(\boldsymbol{R}(\phi_z)\, \boldsymbol{x}_m + \boldsymbol{t}_{2D})|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m).$$

For 3–D objects there are two additional degrees of freedom. They allow an object rotation $\boldsymbol{\phi}_{ext} = (\phi_y, \phi_x)$ perpendicular to the image plane. With the same density model for all possible rotations $\boldsymbol{\phi}_{ext}$ the density parameters are functions of these additional parameters, so that:

$$p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t}) = \prod_{\boldsymbol{x}_m \in A} \mathcal{N}(\boldsymbol{c}(\boldsymbol{R}(\phi_z)\, \boldsymbol{x}_m + \boldsymbol{t}_{2D})|\boldsymbol{\mu}_m(\phi_y, \phi_x), \boldsymbol{\Sigma}_m(\phi_y, \phi_x)).$$

Assuming continuous functions $\boldsymbol{\mu}_m(\phi_y, \phi_x)$, $\boldsymbol{\Sigma}_m(\phi_y, \phi_x)$, they can be rewritten using a basis set for the domain of two-dimensional functions $\{v_r\}_{r=0,\ldots,\infty}$ on the domain of (ϕ_y, ϕ_x) with appropriate coordinates $a_{m,n,r}, b_{m,n,r} \in \mathbb{R}$ $(r = 0, \ldots)$:

$$\mu_{m,n} = \sum_{r=0}^{\infty} a_{m,n,r} v_r, \qquad \sigma_{m,n}^{-2} = \sum_{r=0}^{\infty} b_{m,n,r} v_r.$$

Possible basis functions are $v_r = v_{st}(\phi_y, \phi_x) = \phi_x^s \phi_y^t$ with the enumeration $r = \frac{1}{2}(s+t)(s+t+1)+t$. The functions can be approximated by using only part of the complete basis set $\{v_r\}_{r=0,\ldots,L-1}$. The approximation error can be made as small as possible by choosing L large enough. If ϕ_x is constant, as in our experiments, the v_r are only one-dimensional polynomial basis functions ϕ_y^r.

4.2 Parameter estimation

The model parameters are estimated by a maximum likelihood estimation. Under the assumption of N_ρ independent observations ${}^\rho\boldsymbol{c}_A$ this leads to the estimation

$$\left\{ (\widehat{\boldsymbol{a}}_{m,n}, \widehat{\boldsymbol{b}}_{m,n}) \right\} = \underset{\{(\boldsymbol{a}_{m,n}, \boldsymbol{b}_{m,n})\}}{\operatorname{argmax}} \prod_\rho p({}^\rho\boldsymbol{c}_A | \boldsymbol{x}_m, \{(\boldsymbol{a}_{m,n}, \boldsymbol{b}_{m,n})\}, {}^\rho\boldsymbol{R}, {}^\rho\boldsymbol{t}),$$

with the assumption of known transformation parameters ${}^\rho\boldsymbol{R}, {}^\rho\boldsymbol{t}$ and a predefined number L of basis functions.

The optimization of this function is rather complex. In order to reduce the complexity by providing an analytical solution, $\sigma_{m,n}$ is assumed to be constant.

Solving the equations for the parameters to be estimated results in:

$$\widehat{\boldsymbol{a}}_{m,n} = \boldsymbol{Q}^{-1} \left(\sum_\rho {}^\rho c_{m,n} \boldsymbol{v}({}^\rho\boldsymbol{\phi}_{ext}) \right), \quad \widehat{\sigma}_{m,n} = \frac{1}{N_\rho} \sum_\rho \left({}^\rho c_{m,n} - \widehat{\mu}_{m,n}({}^\rho\boldsymbol{\phi}_{ext}) \right)^2,$$

with

$$\boldsymbol{Q} = \sum_\rho \boldsymbol{v}({}^\rho\boldsymbol{\phi}_{ext}) \boldsymbol{v}^{\mathrm{T}}({}^\rho\boldsymbol{\phi}_{ext}),$$

${}^\rho c_{m,n} = {}^\rho c_n(\boldsymbol{R}\,({}^\rho\phi_z)\,\boldsymbol{x}_m + {}^\rho\boldsymbol{t}_{int})$ and $\boldsymbol{v} = (v_0, \ldots, v_{L-1})^{\mathrm{T}}$.

5 Results

Fig. 3 shows the objects used in this work. The images are 256 pixels in square. Several real images containing one object at different positions have been used for experiments. The localization was performed on two scale levels: $s_0 = 4$, $s_1 = 3$ ($d = 2$ pixels) with resolution $r_s = 0.5d^s$ ($r_{s_0} = 8$, $r_{s_1} = 4$ pixels) and constant $\sigma_{m,n}$. The Gabor filters were calculated for 16 equidistant angles from 0^o to 180^o, resulting in nine-dimensional feature vectors. Only the best localization result of level s_0 was used for further refinement at s_1. The Downhill Simplex algorithm was used for the local parameter search following the global grid search. The

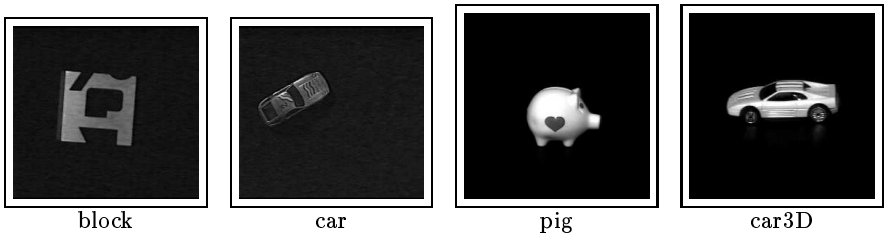

Fig. 3. Examples for objects used for 2–D (left) and 3–D (right) experiments.

computation time on a SGI Impact is about 45 seconds for feature extraction on both scale levels and 30 seconds for localization of object *pig* with its four-dimensional parameter space (level s_0: 100 grid locations, s_1: 400).

Training and test sets are disjoint. For each of the 2–D objects *halfcircle*, *car* and *plug* one image sequence with a complete object rotation in the image plane in 36 equidistant steps was available. The correct object positions were determined manually with two reference points. The average accuracy of the object positions available for training and testing is about half a pixel with respect to translation and half a degree with repect to rotation. Object *block* was available in three such sequences with different lighting conditions. Training images were taken out of two sequences, the rest was used for testing. The correct positions for *block* were determined by the algorithm described in [3]. The sequences of the 3–D objects *pig* and *car3D* are taken from the Columbia Object Image Library (COIL). They consist of images for different object positions of one rotation axis ϕ_y of $\boldsymbol{\phi}_{ext}$ and fixed ϕ_x,ϕ_z,$\boldsymbol{t}$. The range of $\boldsymbol{\phi}_{ext}$ was treated as one-dimensional in the experiments. The range of $\phi_z, \boldsymbol{t}$ was searched completely, resulting in a four-dimensional search. Tables 1 and 2 show the results of the tests. Experiments with a translation error of more than ten image pixels were categorized as failure.

Object	Number		Error			
	Train	Test	Transl. (Pix)		Rot. (^{o})	
			mean	max	mean	max
block	40	66	0.3	1.4	0.5	1.6
halfcircle	18	18	0.8	1.8	0.5	1.6
car	18	18	1.3	3.8	1.7	4.0
plug	18	18	0.9	2.0	0.9	1.7

Table 1. Localization results for 2–D objects.

6 Conclusion

A new approach for object localization using statistical models was presented. The localization scheme works without segmentation of the input images. Gabor filters are used to extract local features. The local features are transformed in order to be rotationally invariant and robust to changes in lighting conditions.

Object	Number		L	Fail	Error					
	Train	Test			Transl. (Pix)		int.Rot. (^{o})		ext.Rot. (^{o})	
					mean	max	mean	max	mean	max
pig	36	36	6	2	1.2	2.9	1.5	4.7	4.8	16
car3D	36	36		4	1.5	7.4	3.0	9.4	6.6	22
pig	36	36	8	0	1.1	2.1	1.5	5.6	3.7	14
car3D	36	36		2	1.7	9.9	2.7	9.4	5.4	16

Table 2. Localization results for 3–D objects.

The positions of the local features with respect to the object provide the information for localization. The object model consists of the distributions of the local feature vectors. Assuming stochastic independence the densities are combined to a complete model for the object. The localization itself is performed on different scale levels to reduce the computational complexity.

Several experiments with real images of 2–D and 3–D objects were carried out. The results show that the approach is capable of localizing objects. Future research will focus on the investigation of alternative and rotationally variant features. Statistical dependence will be considered to a certain degree.

References

1. M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In B. Buxton and R. Cipolla, editors, *Computer Vision — ECCV '96*, volume I of *Lecture Notes in Computer Science*, pages 329–341, Heidelberg, 1996. Springer.
2. J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.
3. J. Hornegger and H. Niemann. Statistical learning, localization, and identification of objects. In ICCV 95 [4], pages 914–919.
4. *Proceedings of the 5^{th} International Conference on Computer Vision (ICCV)*, Boston, June 1995. IEEE Computer Society Press.
5. H. Kollnig and H.-H. Nagel. 3D pose estimation by fitting gradients directly to polyhedral models. In ICCV 95 [4], pages 569–574.
6. V. Kumar and E. S. Manolakos. Unsupervised model–based object recognition by parameter estimation of hierarchical mixtures. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 967–970, Lausanne, Schweiz, September 1996. IEEE Computer Society Press.
7. B. S. Manjunath, C. Shekhar, and R. Chellappa. A new approach to image feature detection with applications. *Pattern Recognition*, 29(4):627–640, 1996.
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. A. Shustorovich. Scale specific and robust edge/line encoding with linear combinations of gabor wavelets. *Pattern Recognition*, 27(5):713–725, 1994.
10. P. Viola and W. Wells III. Alignment by maximization of mutual information. In ICCV 95 [4], pages 16–23.

This article was processed using the LaTeX macro package with LLNCS style