Wavelet Features for Statistical Object Localization Without Segmentation

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 describes a new technique for statistical 3–D object localization. Local feature vectors are extracted for all image positions, in contrast to segmentation in classical schemes. We define a density function for those features and describe a hierarchical pose estimation scheme for the localization of a single object in a scene with arbitrary background. We show how the global pose search on the starting level of the hierarchy can be computed efficiently. The paper compares different wavelet transformations used for feature extraction.

1 Introduction and Motivation

Statistical object modeling without segmentation is motivated by speech understanding. Recognition results were strongly enhanced by the idea of incorporating statistical methods in the recognition task [8] there. Recent publications [2, 9, 5] show, that a statistical framework can also successfully be established in image object recognition.

Most of the publications construct statistical models based on segmentation results. They use the geometric information provided by results, like lines or vertices, as random variables. 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. Segmentation results may be incomplete or located incorrectly.

One way to cope with the problem is to avoid segmentation. Instead of that the gray–level information of an image can be used. [6] describe a recognition method based on singular value decompositions of vector spaces composed of the gray–level data of several images (appearance based modeling). Thereby a large number of images is approximately encoded by a small number of basis images. The projection parameters of an image into this eigenspace can be used for recognition. Maximization of the mutual information between an object model and an object in a scene is a further possibility [9]. A similar technique is described in [4], where the gradient of object models is matched directly to gray-level image sequences of traffic scenes in order to track vehicles. [10] uses the multichannel output of Gabor wavelets to detect and locate 3–D objects in infrared images. [5] 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 referenced papers either use only probabilisticaly chosen model points for matching [9] or use pose restrictions [4] 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. We demonstrate a new way of formulating a statistical model with a functional basis decomposition of probability density parameters. A more detailed description of this model can be found in [7]. Local features are the basis for the recognition process. We compare features derived of different wavelet transformations [1].

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. We assume that the object does not vary in scale.

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) $r_s \in \mathbb{R}^+$ at the locations of rectangular sampling grids. Given an image $f(x, y)$ with $x \in \{0, 1, \ldots, D_x - 1\}$, $y \in \{0, 1, \ldots, D_y - 1\}$ the observed feature values at scale s are denoted by $\boldsymbol{c}_s(x, y) = (c_{s,0}, \ldots, c_{s,N-1})^{\mathrm{T}}$ $(x \in \{0, r_s, \ldots, r_s D_x - 1\},\ y \in \{0, r_s, \ldots, r_s D_y - 1\})$. In the experiments of this paper the features $\boldsymbol{c}_s$ are chosen as the logarithmic coefficients of the scaling functions — that are the low pass coefficients — of a discrete wavelet transform ($N = 1$). We use tensor product wavelets. Only almost symmetric wavelets are chosen to get local features which are robust to object rotations in the image plane.

*The author is member of the Center of Excellence 3-D Image Analysis and Synthesis sponsored by the "Deutsche Forschungsgemeinschaft".

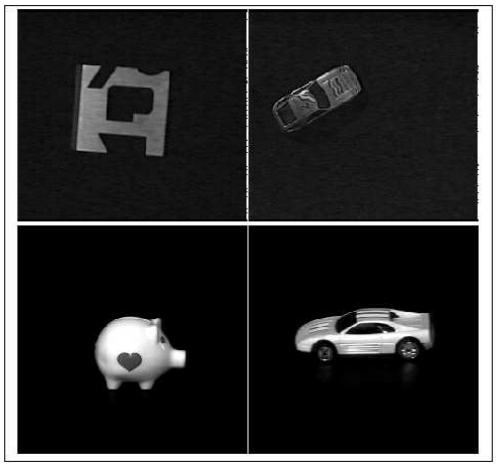

Figure 1: Objects for experiments: *block*, *car*, *pig*, *car3D* (top left to bottom right).

With those features a statistical measure for their probability under the assumption of an object transformation can be defined. 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 (Figure 2). Measures are defined for each scale and the localization is performed for each level successively. Let $\tilde{\boldsymbol{c}}_s$ be the vector of the concatenated feature values detected in an image on scale s, $\boldsymbol{B}_s$ the model parameters of an object class and $\boldsymbol{R}, \boldsymbol{t}$ be the 3–D rotation matrix and translation vector. The rotation $\boldsymbol{R}$ is defined by the rotation angles ϕ_x, ϕ_y and ϕ_z round the x–, y– and z–axis respectively.

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 $(\hat{\boldsymbol{R}}_s, \hat{\boldsymbol{t}}_s) = \text{argmax}_{(\boldsymbol{R},\boldsymbol{t})}\, p(c_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t})$.

3 Statistical model

3.1 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 is omitted.

The model object is covered with a rectangular grid of local feature vectors. The grid resolution is the same as the image resolution on the actual scale. Let $A \subset \mathbb{R}^2$ be a small region (e.g. rectangular) which contains the object projection to the image plane for all possible rotations $\boldsymbol{\phi}_{ext} = (\phi_y, \phi_x)$ outside the image plane. 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 feature vector at location $\boldsymbol{x}$. Assuming the densities $p(\boldsymbol{c}(\boldsymbol{x}_m))$ of the local features as stochastically independent leads to

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

where $\boldsymbol{c}_A$ is the subset of $\boldsymbol{c}$ which is covered by A. The grid positions and the model area A are part of the model parameters $\boldsymbol{B}$.

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. In the case of independence, $\boldsymbol{\Sigma}_m$ is a diagonal matrix $\text{diag}\left(\sigma^2_{m,0}, \ldots, \sigma^2_{m,N-1}\right)$.

The density parameters are a function of the rotation parameters ϕ_y, ϕ_x for 3–D objects, so that:

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

with the 2–D rotation matrix $\boldsymbol{R}\left(\phi_z\right)$ for the rotation and the translation $\boldsymbol{t}_{2D}$ in the image plane. The image feature vectors at the transformed 2–D locations are calculated by linear interpolation. Assuming continuous functions $\boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}_m$ they can be rewritten using a basis set for the domain of two-dimensional functions $\{v_r\}_{r=0,\ldots,\infty}$ with coordinates $a_{m,n,r}, b_{m,n,r} \in \mathbb{R}$ $(r = 0, \ldots)$:

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

The functions are approximated by using only part of the complete basis set $\{v_r\}_{r=0,\ldots,L-1}$. The Taylor decomposition shows, that the approximation error can be made as small as possible by choosing L large enough. With this approximation a fast computation of the density function and a maximum likelihood estimation of the basis coefficients is possible. The estimation results in closed estimation terms if σ_m^{-2} is assumed as constant (see [7]). The value of L is limited mainly by the computation time for the density and the size of the training set for estimation.

3.2 Efficient pose estimation

The pose estimation consists — at least on the roughest resolution level — of a global pose search and succeeding local search. The global pose search evaluates the function $p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t})$ for grid positions covering the possible parameter range. Let $D_{\boldsymbol{R},\boldsymbol{t}}$ denote the number of evaluated transformation parameters. The time complexity of the grid search is then of order $\text{O}(|A|\, LND_{\boldsymbol{R},\boldsymbol{t}})$.

For pose estimation with constant σ^{-2} the function

$$\begin{aligned} &p\left(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t}\right) \\ &= \prod_{m,n} \frac{1}{\sqrt{\left(2\pi\sigma^2_{m,n}\right)}} \exp\left(\frac{-1}{2\sigma^2_{m,n}}(c_n(\boldsymbol{R}\left(\phi_z\right)\boldsymbol{x}_m + \boldsymbol{t}_{2D}) \right. \\ &\qquad \left. - \boldsymbol{a}^{\mathrm{T}}_{m,n}\boldsymbol{v}(\boldsymbol{\phi}_{ext}))^2\right), \end{aligned}$$

has to be maximized with respect to $\boldsymbol{R}, \boldsymbol{t}$.
Applying the logarithm yields the following function to be minimized:

$$h(\boldsymbol{\phi}, \boldsymbol{t}) = \sum_{m,n} \frac{1}{\sigma^2_{m,n}} \left(c_n(\boldsymbol{R}\left(\phi_z\right)\boldsymbol{x}_m + \boldsymbol{t}_{2D}) - \boldsymbol{a}^{\mathrm{T}}_{m,n}\boldsymbol{v}(\boldsymbol{\phi}_{ext})\right)^2.$$

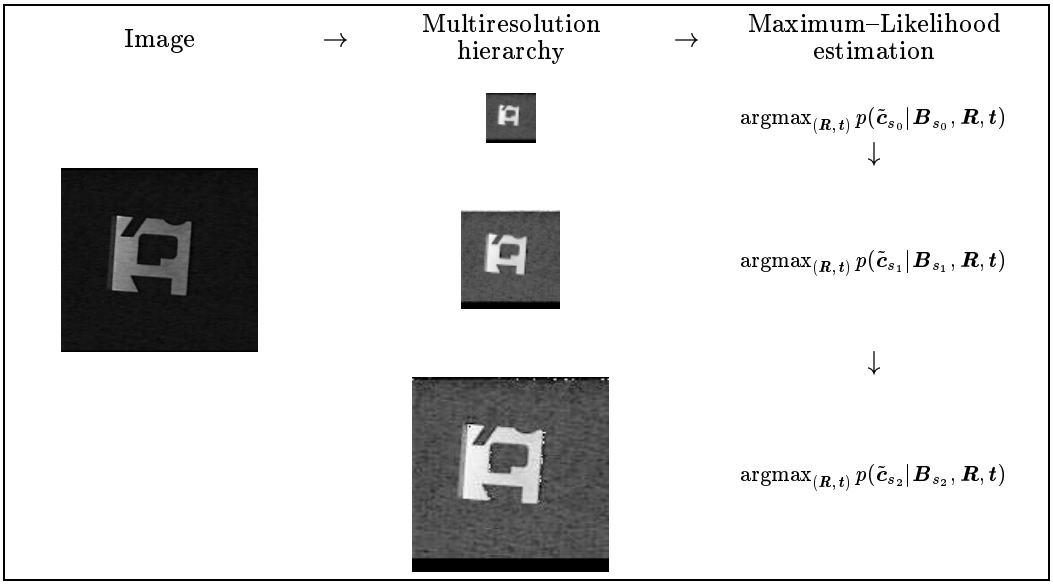

Figure 2: System overview.

With $\boldsymbol{\phi} = (\phi_x, \phi_y, \phi_z)$ and the functions

$$\begin{aligned} h_1(\boldsymbol{\phi},\boldsymbol{t}) &= \sum_{m,n} c_n(\boldsymbol{R}(\phi_z)\,\boldsymbol{x}_m + \boldsymbol{t}_{2D})^2 \sigma_{m,n}^{-2} \\ h_{2,r}(\boldsymbol{\phi},\boldsymbol{t}) &= \sum_{m,n} c_n(\boldsymbol{R}(\phi_z)\,\boldsymbol{x}_m + \boldsymbol{t}_{2D}) \frac{a_{m,n,r}}{\sigma_{m,n}^2} \\ h_3(\boldsymbol{\phi},\boldsymbol{t}) &= \sum_{m,n} \left(\boldsymbol{a}_{m,n}^{\mathrm{T}} \boldsymbol{v}(\boldsymbol{\phi}_{ext})\right)^2 \sigma_{m,n}^{-2}, \end{aligned}$$

the sum $(h_1 - 2\boldsymbol{v}(\boldsymbol{\phi}_{ext})^{\mathrm{T}}\boldsymbol{h}_2 + h_3)(\boldsymbol{\phi},\boldsymbol{t})$ has to be minimized. The global search has to calculate the function values for all grid positions which results in the mentioned complexity, depending on the number of object feature positions.

h_1 and $h_{2,r}$ are of the form

$$\tilde{h}(\boldsymbol{t}_{2D}) = \sum_n \sum_m f_n(\boldsymbol{x}'_m + \boldsymbol{t}_{2D}) w_{m,n},$$

with $\boldsymbol{x}'_m = \boldsymbol{R}(\phi_z)\,\boldsymbol{x}_m$ and for fixed $\boldsymbol{\phi}$. If the evaluation grid $(\boldsymbol{\phi},\boldsymbol{t}_{2D}) \in \{(\boldsymbol{\phi}_{i,j,q}, \boldsymbol{t}_{2D,q,k,l})\}$ for the transformation parameters is chosen as extension of $X' = -\boldsymbol{R}(\phi_{z,q})\,X$ to the possible parameter range, so that

$$\begin{aligned} \boldsymbol{\phi}_{i,j,q} &= (\phi_{x,0} + i\Delta\phi_x, \phi_{y,0} + j\Delta\phi_y, \phi_{z,0} + q\Delta\phi_z) \\ &= (\phi_{x,i}, \phi_{y,j}, \phi_{z,q}) \\ \boldsymbol{t}_{2D,q,k,l} &= -\boldsymbol{R}(\phi_{z,q})\,(t_{x,0} + k\Delta t_x, t_{y,0} + l\Delta t_y)^{\mathrm{T}} \\ &= -\boldsymbol{R}(\phi_{z,q})\,(t'_{x,k}, t'_{y,l})^{\mathrm{T}} \\ X' &= \{-\boldsymbol{R}(\phi_{z,q})\,\boldsymbol{x}_m\} \subset \{\boldsymbol{t}_{2D,q,k,l}\}, \end{aligned}$$

the evaluation of the second sum on X' can be interpreted as convolution. This is, because

$$\begin{aligned} \tilde{h}(\boldsymbol{t}_{2D,q,k,l}) &= \sum_n \sum_m f_n(\boldsymbol{R}(\phi_{z,q})\,\boldsymbol{x}_m + \boldsymbol{t}_{2D,q,k,l}) w_{m,n} \\ &= \sum_n \sum_m f_n(\boldsymbol{R}(\phi_{z,q})\,(\boldsymbol{t}'_{2D,k,l} - \boldsymbol{x}_m)) w_{m,n} \\ &= \sum_n \sum_m \tilde{f}_{q,n}(\boldsymbol{t}'_{2D,k,l} - \boldsymbol{x}_m) w_{m,n} \end{aligned}$$

and

$$\begin{aligned} \sum_m \tilde{f}_{q,n}(\boldsymbol{t}'_{2D,k,l} - \boldsymbol{x}_m) w_{m,n} &= \sum_{\bar{k},\bar{l}} \tilde{f}_{q,n}(\boldsymbol{x}_{k,l} - \boldsymbol{x}_{\bar{k},\bar{l}}) w_{\bar{k},\bar{l},n} \\ &= \sum_{\bar{k},\bar{l}} \tilde{f}_{q,k-\bar{k},l-\bar{l},n} w_{\bar{k},\bar{l},n} \end{aligned}$$

for $\boldsymbol{x}_m = \boldsymbol{x}_{\bar{k},\bar{l}} = \boldsymbol{t}'_{2D,\bar{k},\bar{l}}$.
Using FFT allows the computation of this convolution in $\mathrm{O}(D_t \log(D_{t_x}) \log(D_{t_y}))$ time. The calculation of each of the $L+1$ functions h_1 and $h_{2,r}$ is therefore of complexity $\mathrm{O}(ND_t \log(D_{t_x}) \log(D_{t_y}))$. h_3 has complexity $\mathrm{O}(D_{\boldsymbol{\phi}_{ext}})$. This results in a complexity of order $\mathrm{O}(LND_{\boldsymbol{R},t} \log(D_{t_x}) \log(D_{t_y}))$ for the complete search, where the computation of h out of the simpler functions can be performed very fast.

4 Results

Figure 1 shows the objects used in this work. The images are 256 pixels in square. The localisation was performed on two scale levels s_0, s_1 with resolution $r_{s_0} = 8$ and $r_{s_1} = 4$ pixels and constant $\sigma_{m,n}$. 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 computation time

Wavelet	L	Fail	Error						$q_{\kappa,\kappa'}$
			Transl. (Pix)		int.Rot. (^{o})		ext.Rot. (^{o})		
			mean	max	mean	max	mean	max	
Johnston 8	6	0	0.8	3.4	1.9	5.4	6.3	14	145
Haar		0	0.8	2.1	1.8	3.6	6.2	13	183
Daub. Lapl.		0	0.8	2.1	1.5	3.8	6.4	14	153
Zhu		3	1	2.7	1.9	5.5	6.9	14	84
Johnston 8	8	0	0.7	2.1	1.5	3.2	2.8	7.9	203
Haar		0	0.7	1.8	1.3	3	2.5	7.4	245
Daub. Lapl.		0	0.8	2.1	1.3	3.2	2.6	8.3	201
Zhu		0	1	3.4	1.4	4.8	2.9	10	115

Table 2: Results for object *car3D* (Ω_κ) with reference *pig* ($\Omega_{\kappa'}$)

Wavelet	Fail	Error				$q_{\kappa,\kappa'}$
		Transl. (Pix)		Rot. (^{o})		
		mean	max	mean	max	
Johnston 8	0	0.6	2.8	0.6	1.7	82.3
Haar	0	0.6	2.6	0.7	1.5	81.7
DaubLapl.	0	1.3	4.5	1.3	4.1	79.9
Zhu	12	1.0	3.3	1.1	3.2	42.0

Table 1: Results for object *car* (Ω_κ) with reference *block* ($\Omega_{\kappa'}$)

on a SGI O2 (R10000) is about 20 seconds for feature extraction and localization of object *pig* within the four-dimensional parameter space on both scales.

Training and test sets are disjoint. For the 2–D objects *car* one image sequence with a complete object rotation in the image plane in 36 equidistant steps was available. Object *block* was available in three such sequences with different lighting conditions. 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.

Let $\boldsymbol{B}_\kappa$ denote the model parameters of object class Ω_κ and

$$L_\kappa(\boldsymbol{c}) = \log(\max_{\boldsymbol{R},\boldsymbol{t}} p(\boldsymbol{c}_A|\boldsymbol{B}_\kappa, \boldsymbol{R}, \boldsymbol{t}))$$

the logarithmic density value of the maximum likelihood estimation for an observed feature vector $\boldsymbol{c}$. With a reference object class $\Omega_{\kappa'}$ for a different object and observations $\{^\rho\boldsymbol{c}\}$, $\{^\rho\boldsymbol{c}'\}$ the quality measure $q_{\kappa,\kappa'}$ is defined as

$$q_{\kappa,\kappa'} = \frac{\bar{L}\{^\rho\boldsymbol{c}\} - \bar{L}\{^\rho\boldsymbol{c}'\}}{\sqrt{\mathrm{var}\,(\mathrm{L})\{^\rho\boldsymbol{c}\}}}$$

$$\bar{L} = \frac{1}{N_\rho}\sum_\rho L(^\rho\boldsymbol{c}), \quad \mathrm{var}\,(\mathrm{L}) = \frac{1}{N_\rho}\sum_\rho \left(L(^\rho\boldsymbol{c}) - \bar{L}\right)^2,$$

where $|\{^\rho\boldsymbol{c}\}| = N_\rho$. The measure appraises the selectivity of an objects density function with respect to an alternative reference object and its variance. Table 1 and 2 show experimental results for 2–D and 3–D objects respectively. The Zhu wavelet is a representative of a group of wavelets which are not suitable for object localization. The best wavelets are the Johnston 8 and the simple Haar wavelet.

References

[1] A. Rieder A. K. Louis, P. Maass. *Wavelets.* Teubner, Stuttgart, 1994.

[2] J. Hornegger and H. Niemann. Statistical learning, localization, and identification of objects. In ICCV 95 [3], pages 914–919.

[3] *Fifth International Conference on Computer Vision (ICCV)*, Cambridge, MA, June 1995. IEEE Computer Society Press.

[4] H. Kollnig and H.-H. Nagel. 3D pose estimation by fitting gradients directly to polyhedral models. In ICCV 95 [3], pages 569–574.

[5] 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, Switzerland, September 1996. IEEE Computer Society Press.

[6] 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.

[7] J. Pösl and H. Niemann. Statistical 3–D object localization without segmentation using wavelet analysis. In *Proceedings of the 7th International Conference on Computer Analysis of Images and Patterns (CAIP)*, Kiel, Germany, September 1997, to appear.

[8] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] P. Viola and W. Wells III. Alignment by maximization of mutual information. In ICCV 95 [3], pages 16–23.

[10] X. Wu and B. Bhanu. Gabor wavelets for 3D object recognition. In ICCV 95 [3], pages 537–542.