CLUTTER 16: COMPUTER UNDERSTANDING OF NATURAL SCENES [1]

Version 1

Ulf Grenander, Brown University

Division of Applied Mathematics

Providence, 02912 RI

April, 2002

Computer Vision abounds with techniques for improving images, suppression of noise, enhancing their appearence, coding and transmission and many other similar tasks. This has resulted in the emergence of a powerful imaging technology of great usefulness and applicability.

A related but distinct research area is automated image understanding: computer analysis of images from the real world in terms of significant constituent parts and the interpretation of the images in terms of categories that are meaningful to humans. "Is there an object of interest in the picture ?" Standard pattern recognition methods are clearly insufficient for this purpose; edge detectors, median filters, smoothers and the like can possibly be helpful, but are not powerful enough in order to understand real world images.

Realization of this fact led to the creation of Pattern Theory. Starting from the belief that *natural patterns in all their complexity require complex models* for their analysis, detailed representations of human knowledge to describe and interpret the images acquired from the worlds encountered. Indeed, *we need detailed knowledge representations, tailor made for the particular image ensemble that we want to understand*. Substantial advances have been made in this direction during the last few years. Special image algebras have been developed and used for inference for many fields

(1) microbiology applied to micrographs of mitochondria: Grenander,Miller(1994)

(2) recognition of moving objects using radar sensors: Lanterman (2000)

(3) analysis of images obtained with FLIR: Lanterman, Miller, Snyder(1996)

(4) hand anatomy from visible light images: Grenander, Chow, Christensen,Rabbit,Miller (1996) Keenan (1991)

(5) brain anatomy using magnetic resonance images: Christensen,Rabbit,Miller (1996), Christensen,Rabbit,Miller (1996), Joshi, Miller (2000),Matejic (1998),

[1]Supported by NSF-DMS 00774276

Wang, Joshi, Miller, Csernansky (2001)

(6) pattern analysis in natural scenes: Huang,Lee (1999), Lee, Mumford, Huang (2001), Lee, Pedersen, Mumford (2001), Pedersen, Lee (2001)

In particular (6) is directly related to the subject of the present paper. In Sections 2 and 3 we shall continue these studies for the particular case of *forest scenes*. But first consider two models for scenes in general.

1 Pattern Theoretic Representations of Natural Scenes

. We shall consider two models of natural scenes: The Transported Generator Model, TGM, and the Basic 3D Model, B3M. The first one is given by the scene representation in terms of configurations

$$c = \{s_\nu g_\nu \in \mathcal{C}; \nu = ..., -1, 01, ...\} \tag{1}$$

The generators $g_\nu(\cdot)$ are drawn i.i.d from a generator space G and the s's is a realization of a stochastic point process over a similarity group, $s_\nu \in S$. This is simply a mathematical formalizations of the dictum that the material world is made up of objects and that pattern analysis should be based on this fact. For pattern theoretic notions see Grenander (1993), referred to as GPT.

An important special case of the TGM is when the similarity group S consists of translations $x_\nu = (x_{\nu 1}, x_{\nu 2}) \in \mathbf{R}^2$ in the plane and the configuration space $\mathcal{C}$ is made into an image algebra, $\mathcal{I} = \mathcal{C}_{[R]}$, by the identification rule $R =$ "*add*", see GPT, p. 52, so that

$$I(x) = \sum_{\nu=-\infty}^{\infty} A_\nu g_\nu(x - x_\nu); I \in \mathcal{I}; \quad x \in \mathbf{R}^2; x_\nu = (x_{1\nu}, x_{2\nu}) \in \mathbf{R}^2 \tag{2}$$

where the amplitudes A_ν are iid $N(0, \sigma^2)$ and the generators $g(\cdot)$ describe the shape of the objects in the scene, deterministic or random. The points x_ν form a Poisson process in the plane with an intensity λ.

A theorem proved in Grenander(1999a,1990b) says that images governed by the TGM in (2) has the property that linear operators TI operating on the resulting image algebra satisfy approximately the Bessel K Hypothesis: 1D marginal probability densities are of the form

$$f(x) = \frac{1}{\sqrt{p} \quad \Gamma(p) c^{p/2+1/4}} x^{p-1/2} K_{p-1/2}(\sqrt{2/c}|x|) \tag{3}$$

where K is a Bessel function. It holds for operators T that annihilate constants and under mild regularity conditions on the g's.

Another instance of the TGM is when we use the identification rule $R =$ "*min*" leading to the image algebra with elements

$$I(x) = min[A_\nu g_\nu(x - x_\nu)] \tag{4}$$

This version of the TGM corresponds to the situation when scene images are acquired with a range camera, for example a laser radar.

But the TGM is an extreme simplification of real world scenes. First, it is 2D rather than 3D. Second, related to the first one, it neglects perspective transformations of objects, and third, it uses a stationary stochastic process rather than a space heterogenous one. This makes it even more surprising that the approximation is very precise in most cases as was shown in Srivastava, Liu, Grenander (2001). Indeed, it is seldom one finds such close agreement of theory with data outside of physics and chemistry. This remarkable results makes it plausible that the TGM can be used with some hope of success also for pattern analysis of range (laser radar) images. We shall attempt this in Section 2.

There is a need for firmer support for deriving of algorithms for the recognition of Objects Of Interest (OOI) hidden against a background of natural scene clutter, perhaps hiding part of the OOI. This is offered by the B3M

$$scene = \cup_{\nu} s_{\nu} g_{\nu} \tag{5}$$

with the objects represented by generator templates $g_{\nu} \in G^{\alpha_{\nu}}$; see GPT p.3, and, again, the s's form a stochastic point process over the group S. Here α is a generator index, see GPT p. 19, that divides the generator space into index classes

$$G = \cup_{\alpha} G^{\alpha} \tag{6}$$

The generators here mean the surface of the respective objects, and the index classes could represent different types of objects, trees, buildings, vehicles...

In the case of range pictures it is natural to introduce a 3D polar coordinate system (r, ϕ, ψ) where r means the distance from the camera to a point in space and ϕ is the azimuth angle and ψ the elevation angle so that we have the usual relation

$$x_1 = rcos\phi cos\psi; x_2 = rsin\phi cons\psi; x_3 = rsin\psi \tag{7}$$

A point $x = (x_1, x_2, x_3)$ is transformed into Cartesian coordinates $u = (u_1, u_2)$ in the focal plane U of the camera by a perspective transformation that we shall call T. Hence the range image has the pixel values, in the absence of noise,

$$I(u) = min_{\nu}\{(Ts_{\nu} g^{\alpha_{\nu}})(u)\} \tag{8}$$

This version of the B3M will be used in Section 3.

1.1 Information Status for Image understanding.

It would be a serious mistake to think of scene understanding as a problem with the observer equipped with objective and static knowledge about the world from which the scene is selected. On the contrary, the knowledge, codified into a prior measure, evolves over time and may be different for different observers. The well known visual illusions speak to this; the ambiguities are resolved completely only when additional information about the scene is made available.

Think of a person looking for an OOI in a landscape never before see by him - he will be capable of less powerful inference than some one familiar with the landscape. If we send out a robot to search for vehicles in a forest it is clear that it will perform better if equipped with an adequate map than it would otherwise. This is obvious, but with further reaching implications than may be thought at first glance.

The Hungarian probabilist Alfred Renyi used to emphasize that all probabilities are conditional. We believe in this, and conclude that any realistic approach to Bayesian scene understanding must be based on prior probabilities that mirror the current *information status*. The automatic search in a desert landscape for a vehicle using a photograph taken a day ago will be based on a prior different from the situation with no such photograph, just the knowledge that it is a desert. In the latter case the prior may be a 2D Gaussian stochastic process with parameters characteristic for deserts in that part of the world. In the first the prior may be given via a map computed from the photograph superimposed with another Gaussian processing representing possible changes in the location of the sand dunes during the last day; obviously a situation more favorable for the inference machine.

Other incidentals that could/should influence the choice of prior are, meteorological conditions, observed or predicted, position of sun, type of vegetation, topographical features known or given statistically, presence of artifacts like camouflage, ... For each likely information status we should build knowledge informations of the resulting scenes. This is a tall order, a task that will require mathematical skills and subject matter insight. It should be attempted and it will be attempted!

1.2 Attention Fields.

A consequence of changes in information status is the effect on the prior probability measure. As more and more detailed information becomes available about the scene the prior is strengthened. One important instance of this is the *atterntion field* that codifies information about the OOI. It comes in two forms:

(i) For the TGM the attention field AF is a probability measure over S operating on the OOI in the U image plane; it describes the location of the target as seen by the observer (camera).

(ii) For the B3D representation the attention field AF is a probability measure over the similarity group S operating in the background space X; it describes location and pose of the target in 3D coordinates.

The AF information can be acquired in different ways. It can come from visual inputs, the operator sees what looks as a suspicious object. It can be the result of intelligence reports,perhaps with low credibility and accuracy, concerning the location of the OOI. Or without human intervention, say as the output of a secondary device like FLIR: the OOI emits heat and the FLIR obtains crude estimates of its coordinates. Many other technologies can improve the information status.

2 Pattern Understanding of Forest Scenes.

Let us use the TGM in this section. To create a knowledge representation of forest scenes we first have to decide how detailed it should be. Should it involve individual trees and bushes? Must branches be specified? Should tree type be described in the representation, oak, maple, pine...? It all depends upon the goal we are set for using the representation.

If the goal is to discover man made Objects Of Interest, OOI, say vehicles or buildings, we may not need a very detailed descripton of the forest. On the other hand, if the purpose is to automate the collection of tree type statistics we should include tree type information in the knowledge representation. This is not just segmentation, it involves analysis and understanding of the content in the image.

Let us deal with the second of the two alternatives. With some arbitrariness we have chosen the following four generator indices:

A) ground surface in the foreground, α="ground"

B) sky background, α = "sky"

C) trunk element for individual trees, α = "trunk"

D) close foliage, α = "foliage"

Now we must make the α-definitions precise. Since the TGM is 2D and lives in the image plane, the generators consist of areas in this plane. We therefore introduce *index operators* O^α mapping sub-images $I_X = \{I(x); x \in X\}$, where X is a subset, say a rectangle, in the image plane,

$$O^\alpha : I_X \mapsto \{TRUE, FALSE\} \tag{9}$$

In other words, the index operators are decision functions taking the value TRUE if we decide that the area X is (mainly) covered by a generator $g \in G^\alpha$. It may happen that a set X is classified as more than one α-value. We shall order the way we apply the index operators, one after the other, with the convention that a TRUE value overwrites previous truth values. We have used the order

$ground, foliage, trunk, sky$.

In tabular form:

Generator Index Classes		
Generator Index	Informal Definition	Formal Definition
α = ground	Smooth Surface	{X:O^{ground}[I]=1}
α = foliage	Irregular Surface	{X:$O^{foliage}$ [I]=1}
α = trunk	Narrow Vertical Surface	{X: O^{trunk}[I]=1}
α = sky	Infinite Distance	{X: O^{sky}[I]=1}

2.1 Index operator for"ground"

A ground area is usually fairly flat except for the presence of boulders and low vegetation like bushes. We shall formalize this by asking that the gradient be small

$$O^{ground}[I(X)] = TRUE \leftrightarrow \|grad(I)(x)\| < c_1 = 0; x \in X \tag{10}$$

2.2 Index operator for"foliage"

For foliage on the other hand, the leaves give rise to local variation but this variation is moderate as long as the leaves belong to the same tree. Hence we introduce

$$O^{foliage}[I(X)] = TRUE \leftrightarrow c_2 < \|grad(I)(x)\| < c_3 = 0; x \in X \tag{11}$$

2.3 Index operator for"trunk"

Trunk areas are narrow and vertical with small local variation compared to their immediate environment. In Figure 1 the narrow rectangle Xcenter should correspond to a part of a trunk, it is surrounded by two somewhat larger rectangles Xleft and Xright with some overlap. Compute the variances of the pixel values belonging to these three sub-images, call them $Var_{left}, Var_{center}, Var_{right}$ and define the index operator

$$O^{trunk}[I(X)] = TRUE \leftrightarrow Var_{left} > c_4 Var_{center} \quad and$$

$$Var_{right} > c_4 Var_{center}; x \in X \tag{12}$$

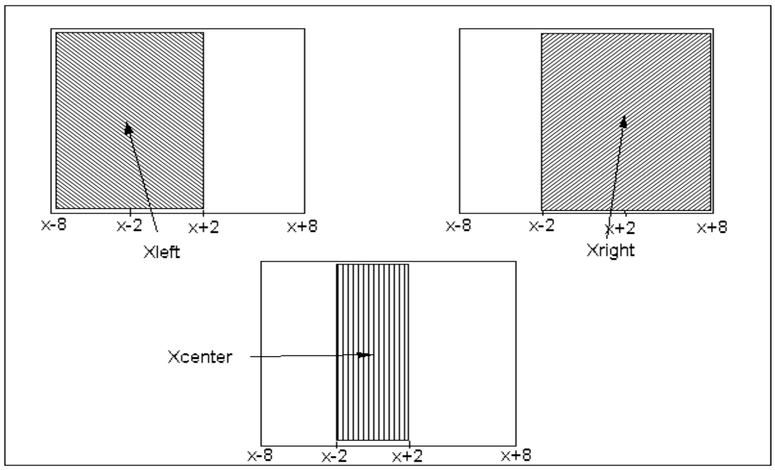


Figure 1

2.4 Index operator for"sky"

This is the easiest generator class to detect. Indeed, sky means infinite distance, far away, or rather the largest distance that the laser radar can register. For the camera used, this is coded as $I(x) = 0$. Hence we can simply define

$$O^{sky}[I(X)] = TRUE \leftrightarrow I(x) = 0; x \in X \tag{13}$$

2.5 Application to Range Images

. Applying this to laser radar images from Lee-Huang(2000):Brown Image Data Base, we get the following analysis with the meanings displayed graphically.

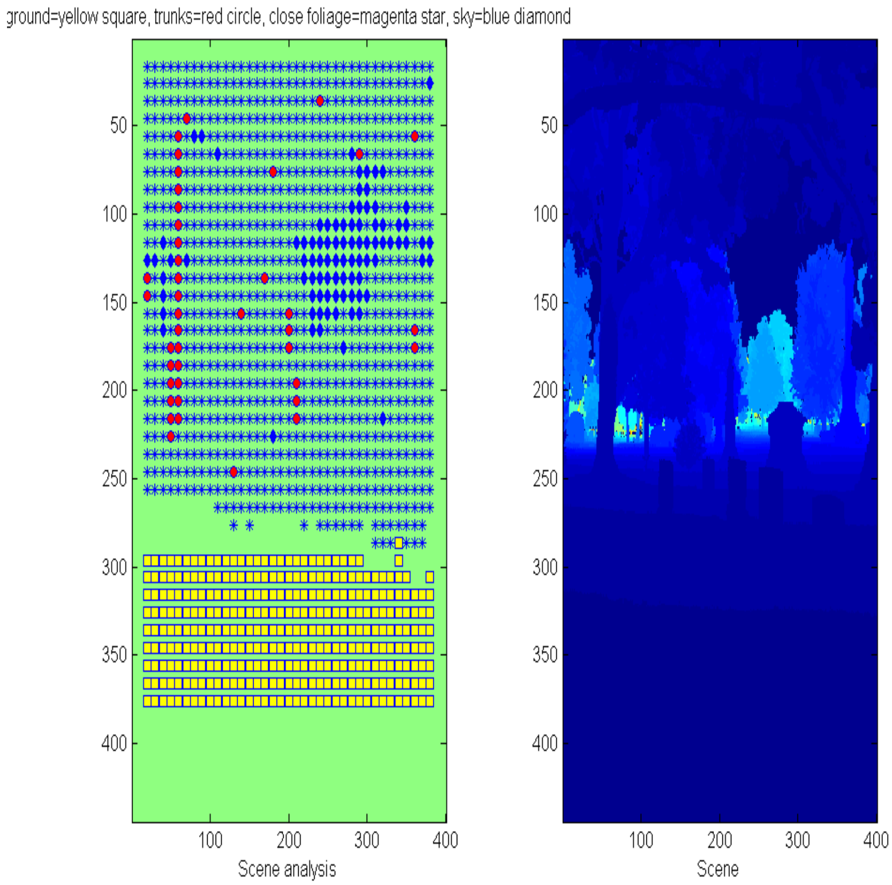


Figure 2

One sees the occurrence of blue diamonds, "sky" and two trunks, some smaller trunk elements too, and a lot of foliage. At the bottom of the figure is the ground, separated from the foliage by pixels that could not be understood by the algorithm. Note that the tombstones in the observed scene have been interpreted as foliage: the present knowledge representation does not "understand" tomb stones. We shall study the understanding of such OOI's in Section 3. In Figure

3 the dominating feature of the analysis is the occurrence of two big trunks.

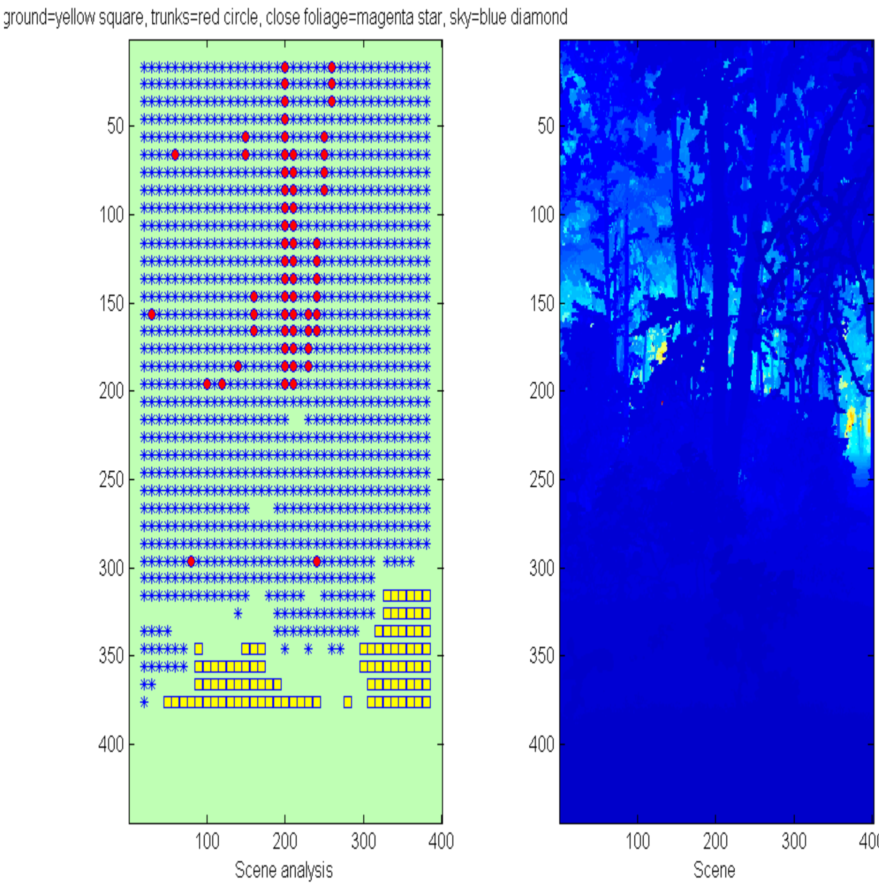


Figure 3

The analysis of Figure 4 also is dominated by a trunk.

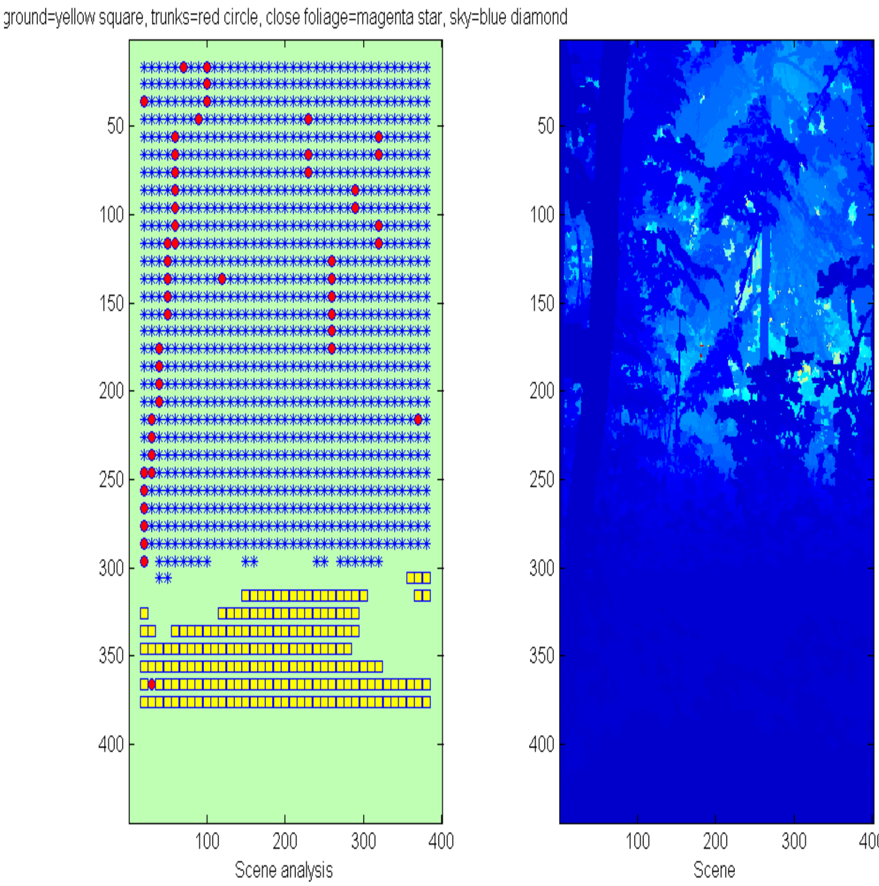


Figure 4

The analysis of Figure 5 has a sky element again.

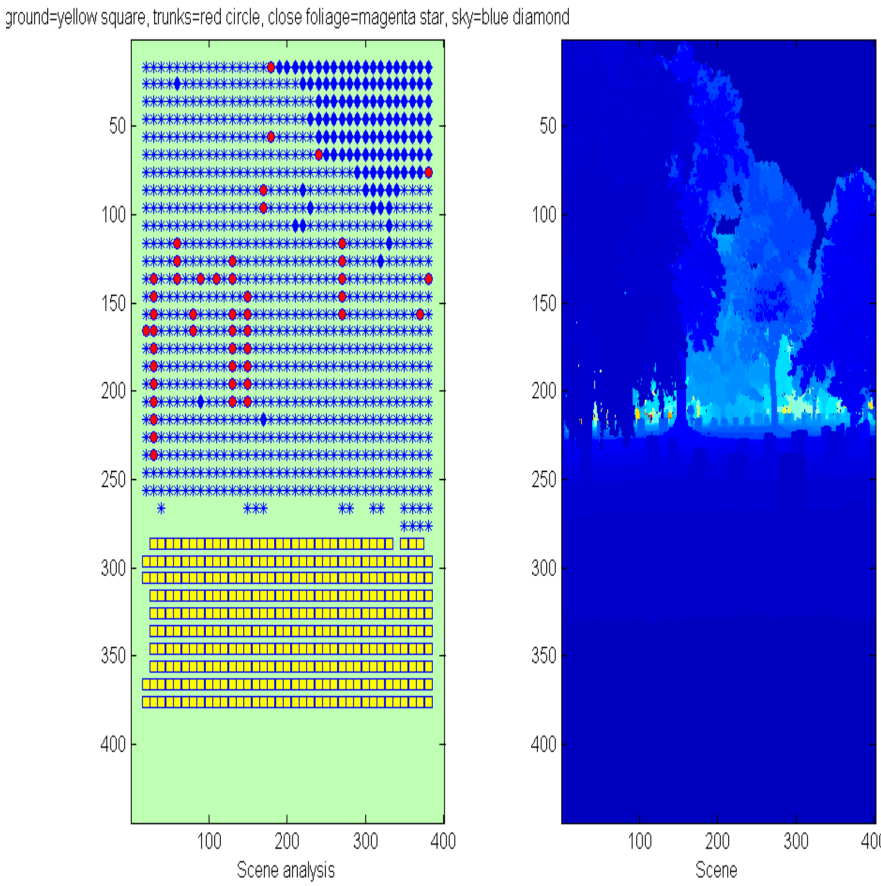


Figure 5

It seems that this automated image understanding performs fairly well considering the primitive form of the index operators that we have used. As we have argued repeatedly before, a major task in the creation of tools for the automate understanding of image ensembles is to *build specific, tailor made knowledge representations in pattern theoretic form for the image ensembles in question.* This is a real challenge, one that has been successfully implemented only for a few scene ensembles, but is a *sine qua non* for any serious attempt to automate scene understanding.

3 Recognizing Objects Of Interest with Forest Clutter Background.

When we turn to the problem of automated recognition of OOI's against a background of clutter, say a forest scene, the role of the background changes. In the previous section the background, the forest, was of primary interest, but now it plays the role of nuisance parameter in statistical parlance. We are now not after an analysis into foliage, trunks..., but cannot neglect the clutter as irrelevant. It has long been know that the randomness of clutter is far from simple white noise. Further, it cannot be modelled as a stationary Gaussian process in the plane. Indeed the validity of the Bessel K hypothesis contradicts such a model from the very beginning. We shall use the TGM for representing the background clutter - the secondary element of the images - and the more detailed B3M for desribing the OOI's - the primary element. For the OOI we choose tanks; we happen to have available a template library in the form of CAD drawings with rotation angles equal to 0,5,10,15,20... degrees. Since we position them on the ground we will have $s3 = 0$, so that this coordinate can be left out in the computations.

Hence the image representation will take the form

$$I(u) = min\{I_{clutter}(u), Tsg^{OOI}(u); u \in \quad image \quad plane \quad U\} + e(u) \qquad (14)$$

for the time being with only a single α-value for the OOI, and $e(\cdot)$ standing for the camera noise. For laser radars the noise level is low, perhaps white noise $e = N(0, \sigma^2), \sigma^2 << 1$. Further the clutter part of the image will be represented as

$$I_{clutter} = \sum_{\nu} A_\nu g(u - u_\nu) \qquad (15)$$

We know some approximate probabilistic properties of such representations, Grenander(1999 a,b).

3.1 Half Bayesian Inference for Range Images

Denote the observed range image by $I^{\mathcal{D}}(u)$. If the noise level is low $I^{\mathcal{D}}(u) \approx I(u)$ for u-points not obscured by the OOI. With the usual Bayesian reasoning applied to the present similarity group we shall, however, not assume any prior for the background; that will be done in the next section. Let us deal with a posterior density $p(s|I^{\mathcal{D}}(\cdot)$ of the form

$$p(s|I^{\mathcal{D}}(\cdot)) \propto \pi(s) exp\{-\frac{1}{2\sigma^2}\|I^{\mathcal{D}} - min(I^{\mathcal{D}}, Tsg^{OOI})\|^2 \qquad (16)$$

with a prior density $\pi(\cdot)$ on the similarity group S expressing the current information status. We admit that it is not clear to what degree this approximation is adequate for inference.

We shall choose the attention field AF as a rectangle in the $(s1, s2)$-plane with the center at the *pointer* $(s1^{AF}, s2^F)$ and with some width $(w1^{AF}, w2^{AF})$, preferably biased bised as mentioned below. Let us try a prior of the form

$$\pi(s) \propto exp[-\frac{(s1-ponter(1))^2}{\sigma_1^2} - \frac{(s2-ponter(2))^2}{\sigma_2^2} - \frac{(s3-ponter(3))^2}{\sigma_3^2}; (s1,s2) \in AF; 0 \quad else \tag{17}$$

not depending on $s4$, so that the three first components are independent and Gausian when restricted to AF, while the fourth one is uniform on $\mathbf{T}$.

Now search for the MAP estimator. Starting at $s = pointer$ and using the Nelder-Mead algorithm, see Nelder,Mead (1965), for function minimization, we will get convergence to a local mimimum, hopefully close to the true s-value, at least if *width* is small enough. But the behavior is a bit puzzling. Sometimes it works well, sometimes not. Why is this so?

One reason is the occurrence of the *min* operation in (16), typical for range cameras. If we let $s2$, the distance away from the camera, be large, the OOI will eventually be hidden behind the clutter, so that $ID - min(ID, Tsg^{OOI}) \equiv 0$ and only the π factor will play a role; the information in the observed image will have been wasted, and we will get misleading inferences.

REMARK 1. It follows that if we use straight maximum likelihood estimation without constraints, *the ML estimator is not consistent.*

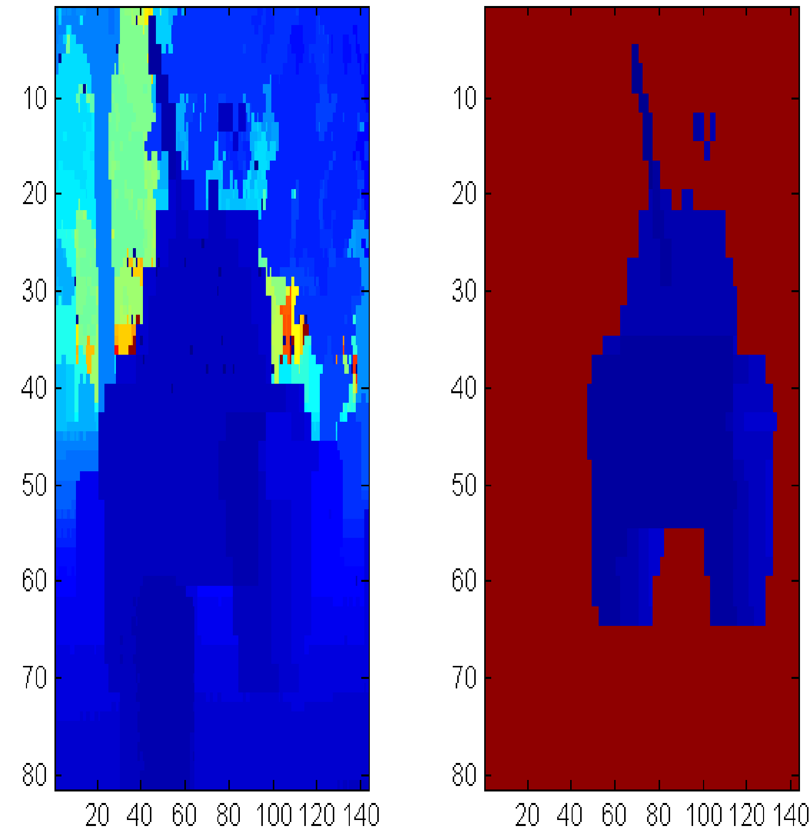
Further, if most or all of the OOI is hidden by the clutter we cannot expect good inference. On the other hand, if only part of the OOI is hidden, the inference algorithm should work. This in contrast to methods that are not designed to take care of obscuration effects, for example simple correlation detectors.

REMARK 2. A technical difficulty is that during the computation all the components of the similarity group element are discrete. This is not serious however; it can be dealt with.

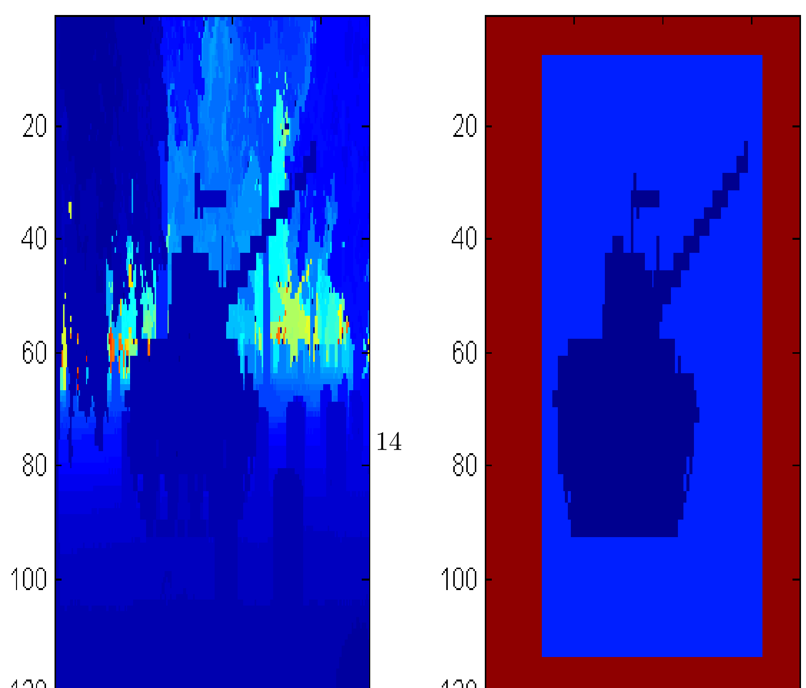
In order to make the inference algorithm work we should, as mentioned above, *introduce a bias in the choice of the pointer* favoring smaller values of $s2$. If we do this the inferences are often quite good. We have found it better, however, to search discretely in the domain of the AP on a fairly coarse grid. More precisely, we first search the translation components and the, keeping the translation fixed, search over $s4$. Although we have not looked for fast algoritms this one seems adequate considering that it is quite crude.

We now apply the algorithm to the range images and display the result by showing a subset of the image containing the OOI and also the *same* subset with the result of the algorithm. In each case the AF was chosen so that it covered the OOI but with some shift in the "pointer" away from the center of the object; this to correspond to mistakes in the information status. We get

Figure 6



The inference looks good. Also for Figure 7



The same is true for the partly hidden OOI in Figure 8

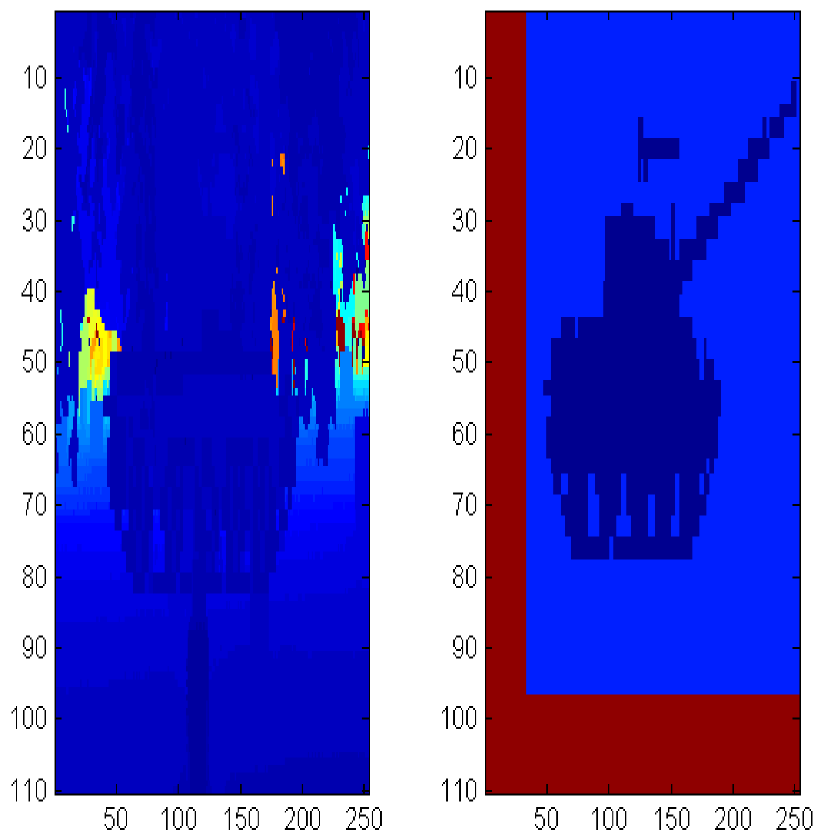


Figure 8

but the algorithm fails of course in the case of the wholly hidden OOI in

Figure 9 and does not discover the OOI

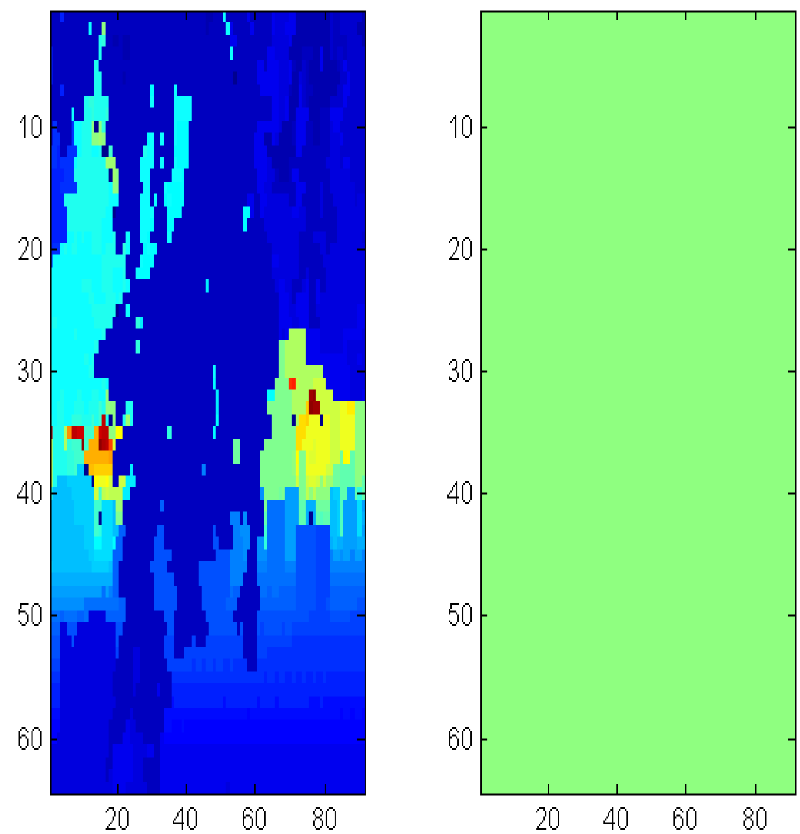


Figure 9

Once the OOI has been found in terms of location and pose the way we have described, it should be possible to implement a scheme of understanding by running the algorithm for several different template libraries, to see which type of OOI we have in the observed image. Lacking an alternative template library we have not implemented this, and just note that the recognition must take into account the variability of the window of decision (the subimage used for recognition) due to variations in perspective when $s2$ varies.

The performance of this algorithm seems acceptable in these cases, but we do not know if this is so in general. It is clear that a theoretical study of its asymptotic properties is needed.

3.2 Fully Bayesian Inference for Range Images.

While the algorithm performed fairly well it is likely that it could be improved by including a prior P over the image algebra $\mathcal{I}$, so that we would adopt a fully Bayesian point of view, assuming of course that we have access to a reliable prior $\mathcal{I}$. Then (16) changes into

$$p(s, I|I^{\mathcal{D}}(\cdot) \propto P(I)\pi(s)exp\{-\frac{1}{2\sigma^2}\|I^{\mathcal{D}} - min(I^{\mathcal{D}}, Tsg^{OOI})\|^2 \tag{18}$$

with the prior density $P(I); I \in \mathcal{I}$. Using results on the Bessel K hypothesis we suggest the form

$$P(I) \propto \prod_{x \in AF} b[I(x_1+1, x_2) - I(x_1, x_2); p_{hor}, c_{hor}] \times b[I(x_, x_2+1) - I(x_1, x_2); p_{vert}, c_{vert}] \tag{19}$$

with p's and c's meaning the parameters in the Bessel K formalism. To make (19) meaningful we must also specify the boundary conditions on AF. A special case of some interest is $p = 1$, the double exponential density

$$b(y; 1, c) \propto exp[-y/c] \tag{20}$$

The resulting, fully Bayesian, recognition algorithm would then compute the minimum of $p(s, I|I^{\mathcal{D}}(\cdot)$ over $S \times \mathcal{I}$ say starting with $I = ID$, and with s corresponding to the value of "pointer". It is easy to write down the Euler equations for I and s, but we have not yet implemented this procedure in code, so we do not yet know if it performs beter than the one in Section 3.1.

4 Selected References

G. Christensen, R. Rabbit, M. Miller(1996): Deformable Templates Using Large Deformation Kinematics. IEEE Transactions on Image Processing, 5(10)

G. E. Christensen, S. C. Joshi, M. I. Miller(1997): Volumetric Transformation of Brain Anatomy IEEE Transactions on Medical Imaging, vol. 16, pp. 864-877

B. Gidas, A. Lee, D. Mumford: Mathematical Models of Generic Images

U. Grenander, A. Srivastava, M. Miller(2000): Asymptotic Performance Analysis of Bayesian Object Recognition, IEEE Trans. on Information Theory, vol. 46, no. 4, pages 1658-1665

U. Grenander, Y.S. Chow, D. Keenan (1991); HANDS, A Pattern Theoretic Study of Biological Shapes, Springe-Verlag

U. Grenander, M. I. Miller(1994):Representation of Knowledge In Complex Systems (with discussion section)," Journal of the Royal Statitical Society, vol.

56, No. 4, pp. 569-603

U. Grenander(1993): General Pattern Theory, Oxford University Press

U. Grenander(1999a): Clutter 5, Tech.Rep.,www.dam.brown.edu/ptg

U. Grenander(1999b): Clutter 6, Tech.Rep.,www.dam.brown.edu/ptg

J. Huang and A.B. Lee (1999): Range Image Data Base, Metric Pattern Theory Collaborative, Brown University, Providence RI.

S. Joshi and M. I. Miller(2000): Landmark Matching Via Large Deformation Diffeomorphisms IEEE Transactions on Image Processing, Vol. 9, No. 8, pages 1357-1370

A. Lanterman(2000): Air to Ground ATT, www.cis.jhu.edu/wu-research/air-to-ground.html

A Lanterman, M.I. Miller, D.L. Snyder(1996): Representations of thermodynamic variability in the automated understanding of FLIR scenes," Automatic Object Recognition VI, Proc. SPIE VOL. 2756, 26-37

A. Lee, J. Huang(2000): Brown Range Image Database, www.dam.brown.edu/ptg/brid

A. Lee, D. Mumford, J. Huang (2001): "Occlusion Models for Natural Images: A Statistical Study of a Scale-Invariant Dead Leaves Model", International Journal of Computer Vision, 41(1/2): 35-59.

A. Lee, K. Pedersen, D. Mumford (2001) "The Complex Statistics of High-Contrast Patches in Natural Images", In Proc. of IEEE Workshop on Statistical and Computational Theories of Vision, ICCV 2001, Vancouver, CA,

L. Matejic:(1998): Testing for brain anomalies: A hippocampus study. Journal of Applied Statistics, 25(5):593-600

M.I. Miller, S.C. Joshi, D.R. Maffitt, J.G. McNally, U. Grenander(1994):Membranes, Mitochondria, and Amoebae: 1, 2, and 3 Dimensional Shape Models, in Advances in Applied Statistics, (ed. K. Mardia), Oxford, England: Carfax, vol. II, pp. 141-163

D. Mumford(1966): Pattern theory: a unifying perspective, in "Proc. 1st European Congress of Mathematics", Birkhauser-Boston, 1994, and in revised form in "Perception as Bayesian Inference", edited by D.Knill and W.Richards, Cambridge Univ. Press.

J.A. Nelder, R. Mead(1965): A simplex method for function minimization, J. Comp., 7, 308-313

K.S. Pedersen, A.B. Lee (2001): Toward a Full Probability Model of Edges in Natural Images. APPTS Report 02-1, DAM Brown University, to appear in the ECCV Proceedings, Copenhagen, May 28-31, 2002

A. Srivastava, X. Liu and U. Grenander. Universal analytical forms for modeling image probabilities,,www.dam.brown.edu/ptg

L. Wang, S. Joshi, M. Miller, J. Csernansky(2001):, "Statistical Analysis of

Hippocampal Asymmetry in Schizophrenia," NeuroImage, 14, 531-545