Strategies for Seeing

Ulf Grenander

Abstract. We shall study the mathematical basis for computer vision using ideas from pattern theory. Starting from some general principles for vision several strategies for seeing will be derived and implemented by computer code. Using the code computer experiments have been carried out in order to examine the performance of the resulting inference engines for vision.

1991 Mathematics Subject Classification: 62P
Keywords and Phrases: pattern theory, computer vision

The mathematics of vision is not well understood. The human visual system is an awesome inference engine of unparalleled power, but its working remains a mystery in spite of great advances in the study of vision in recent years: much is known about its detailed functioning on the physiological level but theories proposed about its overall logical architecture are still tentative.

After the appearence of David Marr's seminal work, Marr (1982), many researchers in vision have adopted his view that seeing should be treated as a computational activity, where 'computational' is understood in a wide sense, more general than von Neumann architecture or Turing machines. We adhere to this view although we do not insist on his feed-forward paradigm. Therefore we believe that there should be a mathematical theory of vision underlying the visual computing and that machine vision would be aided by such a theory.

Another difference to Marr's approach is that we shall emphasize the primacy of analysis of the environment: this is needed for the understanding of the 'why' and 'how' of the algorithms that are realized through the sensory processing. An early proponent of this research strategy was Gibson with his 'ecological psychology', Gibson (1979).

To analyze the environment, the scene ensemble to be encountered by the visual system, we shall apply ideas from pattern theory and will use methods from this discipline as presented in Grenander (1993). A similar approach to vision, but oriented toward human rather than machine vision, has been outlined in Mumford (1994), (1996).

The vision strategies will be reductionist in the sense that they will be *derived* from general and mathematically articulated prinicples in contrast to being based on ad hoc devices. To achieve this the starting point will be the mathematical representation of the image algebra of the likely scenes. Different representations will lead to different strategies for seeing. Several strategies have been derived and implemented computationally. We do not attribute much significance to the algorithms themselves, since they are based on quite simple minded representations, but more to the way they are derived from first principles.

1. Principles for vision.

1. *To be able to see it is necessary to know what one is looking for.* In other words, the system must be equipped with knowledge about scenes that are likely to be encountered and be based on an explicitly formulated purpose. It is therefore the first task for the system designer to express such knowledge in a form that is sufficiently precise for the software development. In a biological system such knowledge may have been created and stored during evolution, but we shall only be concerned with computer vision in the following. The system must also possess the ability to handle scenes it is not expecting, send warning signals and be honest enough to admit ignorance in doubtful situations.

2. *Different scene types and different sensors will require different strategies of vision.* To ask for a universal vision system, a system that is able to see and interpret anything, any electro-magnetic radiation emanating from completely arbitrary scenes, is a hopeless task. Instead of searching for such a chimera we shall narrow down and specify the ensemble of scenes that the system is intended for. We do not believe there is any universal representation valid for all scene/sensor combinations. Therefore the representations must be tailored to the particular scene types.

3.*Knowledge about the image esemble should be represented by logical structures formulated so precisely that they can serve as a basis for computing.* The representations shall be *compositional* in the sense that scenes are built from geometric objects, generators, that are combined together according to rules that may be deterministic or stochastic. They shall be *transformational* in that generators are themselves obtained from prototypes,*templates*, that are modified by transformations that play the role of generalizations.

It is clearly impossible to store all expected scenes in memory: this is avoided by the compositional/transformational scheme. Compare Chomskyan linguistics.

4.*The transformations shall form groups, arranged in a cascade that starts with solid, often low-dimensional, transformations and ends with diffeomorphisms.* The cascade will typically begin with translation, rotation, and perhaps scaling groups, whose semi-direct product forms a low dimensional group S_{solid}, but greater flexibility is needed to get enough generative power to deal with complex image ensembles of high variability and that will be supplied by the full diffeomorphic group S_{diff} or one of its high-dimensional sub-groups. The idea of group cascades has been examined in Matejic (1996). To represent abnormal variability it may be necessary to extend the transformations by giving up the group property, but this will not be explored here.

5. *The occurence of templates in the scene is controlled by probabilities, and deformations of the templates will be controlled by other probability measures on the groups; these measures evaluate how likely are the occurrences of various transformation of the templates.* Consider the set $C = S_{diff}/S_{solid}$ of right (or left) cosets of the sub-group S_{solid} in the full group S_{diff}.

The elements in C represent shape changes while S_{rigid} describes the less drastic transformations that moves sets around etc. The cosets can carry vital information while the elements of S_{solid} often play the role of nuisance parameters in the statistical sense of this term.

6. *The mechanism T that maps a scene into sensory entities shall be explicitly defined.* In general we shall let $\mathcal{T}$, the range of the T's, consist of arrays, not necessarily rectangular, with scalar entries and of fixed shape.

7. *The T transformation can be controlled by the system.* This allows the system to concentrate its attention on a detail of the scene, to direct its sensor(s) to point in a new direction or vary the focal length. In animal vision this corresponds to focussing the *fovea* and it also enables the vision system to function at different scales.

8. *The control of T is governed by an attention function A that attributes different weights to different parts of the observed image $I^{\mathcal{D}}$.* We should think of A as a real valued function of sub-images of $I^{\mathcal{D}}$ that takes real values, $A : 2^{I^{\mathcal{D}}} \to \mathbf{R}$. The attention function formalizes the purpose(s) of the vision engine.

9. *The saccadic search will be controlled by covariants w.r.t. the solid groups.* This will suggest plausible candidates for the generators that make up the true scene.

10. *For fixed T visual understanding will be attempted by an inference engine that selects plausible generators and elements from the groups that deform these generators.* In this way local decisions are made sequentially, forming, accepting or rejecting hypotheses. The selection may be deterministic, say maximizing some estimation criterion, or have random elements, as in simulating a posterior distribution. The visual understanding shall result in a structured description of the scene that can be used for decision making.

11. *The saccadic search is intended to reduce global inference problems to local ones.* The saccads should give rough estimates of the true group elements; the estimates will then be refined by applying the local group operations applying the diffeomorpic deformations.

12. *Once a ROI (Region Of Interest) has been analyzed the attention function is examined again to find other possible ROIs.* If the saccads result in more than one ROI they are all analyzed in the same way until the attention function points to no more ROI.

13. *The noise in the system is represented by a stochastic process N operating on the array outputted by the sensor transformation T.* Note that this randomness

is essentially different from the one governing the variability of the scenes. The latter is inherent in the vision setup, while the first can differ from sensor to sensor.

2. Mathematical Formalization.

2.1. Let us now express the principles mathematically and concretize to specific choices of assumptions that have been used in a series of computer experiments. But first let us explain what we mean by a solid group, or rather solid group action, in a general context, Consider a configuration of generators g_i coupled by the connector σ

$$c \ = \ \sigma(g_1, g_2, \ldots g_n) \ = \ \cup_k c^k$$

where the sub-configurations c^k are the connected (w.r.t. the neighborhood system induced by the graph σ) components of c. Then we shall define a general solid transformation to be of the form

$$c \ \mapsto \ \cup_k s^k c^k; s^k \in S_{solid}$$

so that each connected component is transformed separately and with the same group element for all the generators in the component and with the semi-direct product $S_{solid} = SL(d) \propto \mathbf{R}(d)$ of the special linear group with the translation group in d dimensions.

To formalize principles 1 - 3 let the generators form a space partitioned into the subsets G^α

$$G \ = \ \cup_\alpha G^\alpha$$

where α denotes the object type.

Principle 4 will be realized by choosing some of the sub-groups of S_{solid} and S_{diffeo}.

The purpose of the cascade is to allow large deformations, which is not possible with the single group elastic model, but without the large computing effort needed for the fluids model, see Christensen, Rabbit, Miller (1993).

Principle 5 will be implemented by introducing probability measure on the groups which is straightforward for the solid ones since they are low-dimensional. For S_{diffeo} (discretized approximation) on the other hand we induce a probability measure via the stochastic difference equation

$$(Ls)(x) \ = \ e(x); x \in X$$

for the displacement field $s(x) = (s_1(x), s_2(x))$ and $e(x)$ is a stochastic field; the group action is $x \mapsto x + s(x)$. Let us choose basis functions for S_{diffeo} (discretized to a lattice $\mathbf{Z}_{l_1 \times l_2}$) as the eigen functions of L as in Grenander (1993), p. 523,

$$\phi_{\mu\nu}(x) \ = \ sin(\frac{\pi x_1 \mu}{l_1}) sin(\frac{\pi x_2 \nu}{l_2}); x = (x_1, x_2) \in [1, l_1] \times [1, l_2]$$

with $\mu, \nu = 1, 2, \dots r$, where the choice of r depends on the resolution of the sensor. Then we can expand the displacement fields

$$s_1(x) = \sum_{\mu=1}^{r} \sum_{\nu=1}^{r} t_{1\mu\nu} \phi_{\mu\nu}(x)$$

$$s_2(x) = \sum_{\mu=1}^{r} \sum_{\nu=1}^{r} t_{2\mu\nu} \phi_{\mu\nu}(x)$$

and we combine the Fourier coefficients into two matrices

$$t_1 = (t_{1\mu\nu}; \mu, \nu = 1, 2, \dots r)$$

$$t_2 = (t_{2\mu\nu}; \mu, \nu = 1, 2, \dots r)$$

We shall assume that for each generator index α the set G^α can be generated by applying S_{diffeo} to a single template g^α_{temp} so that

$$G^\alpha = S_{diffeo} g^\alpha_{temp}$$

In pattern theoretic terminology G^α then forms a pattern, actually a *finest pattern*, see Grenander (1993) p. 55-56. Then (G^α, S_{diffeo}) forms a homogeneous space.

For principle 6 we shall allow the range $\mathcal{T}$ of the T-transformation to be quite different from the scene that is being captured. For example, the output of a radar with a cross array of antennas will consist of two vectors with complex entries, superficially completely different from the target/background configuration. Or, the sinogram in a CAT scan which is quite different from the organ scanned.

Principles 7,8 will be realized by attention functions that will formalize the purpose of the system. For example, it could give great weight to regions close to the sensor, or to regions with high optical activity, or to objects of particular shape or texture. The function will generate saccadic movement of the fovea and/or the sensor(s).

For principle 9 we shall use classical covariants. Say that the intensity functions $I(\cdot)$ are continuous with compact support. For example, dealing with the translation group in the plane we use the 2-vector valued covariant

$$\phi^1(I) = m = 1/J \int\int (x_1, x_2) I(x_1, x_2) dx_1 dx_2$$

with

$$J = \int\int I(x_1, x_2) dx_1 dx_2$$

For $SO(2)$ we calculate the moment matrix

$$R = 1/J \int\int (x - m)(x - m)^T I(x) dx_1 dx_2$$

and diagonalize it $R = O^T DO$ and put

$$\phi^2(I) \ = \ O$$

Note however that this definition needs a further qualification in order to be unique. First, we should choose the orthogonal matrix O so that $det(O) > 0$ since we are dealing with the *special* orthogonal group $\mathbf{SO}(2)$. Second, we should select O so that its first column equals the eigen vector of R corresponding to the largest eigen vallue. The sign of the eigen vector is arbitrary so that this leads to an ambiguity that must be kept in mind when developing the code. Third, if the two eigen values coincide, typical for symmetric objects, we get more ambiguity and the covariant must be augmented with further information.

For the uniform scaling group in the plane we can use the scalar covariant

$$\phi^3(I) \ = \ 1/J \int \int \|x\| I(x) dx_1 dx_2$$

The use of saccadic search has split the global inference problem into several local ones in which we can let the inference engine look just for a local optimum, principles 10, 11.

Of course the whole group can push templates outside the total (bounded) region $\mathbf{Z}_{(l_1,l_2)}$, so that search should be limited to the latter region unless the sensor is re-directed to some other region. The saccadic search will lead to one ROI after another, point 12, until the remaining attention values are suffiently small; then the inference engine stops and outputs a structured description of the scene.

For principle 13 let us assume that the noise process of the system forms a stationary process in the plane, for example the Gaussian one with the non-singular covariance operator Cov. Then the likelihood function will be proportional to

$$L \ = \ exp - \frac{1}{2\sigma^2} \| I^{\mathcal{D}} \ - \ TsI_{temp} \|^2_{Cov^{-1}}$$

with the norm associated with the kernel Cov^{-1}. Introducing the positive definite square root M

$$M \ = \ +\sqrt{Cov^{-1}}$$

we can write the likelihood function in terms of the standard l_2-norm

$$L \ = \ exp - \frac{1}{2\sigma^2} \| MI^{\mathcal{D}} \ - \ MTsI_{temp} \|^2 =$$

$$= exp - E_{likelihood}$$

where $E_{likelihood}$ is the likelihood energy.

In a similar fashion we are led to prior probability measures on each group in the cascade. For the S_{diffeo}, for example, we have used the expression

$$E_{prior}(s) \ = \ 1/2\sigma^2 \sum_{\mu=1}^{l_1} \sum_{\nu=1}^{l_2} [l_1 \mu^2 t^2_{1\mu\nu} \ + \ l_2 \nu^2 t^2_{2\mu\nu}]$$

involving the t-matrices introduced earlier and with some scaling constant σ^2. We can then apply Markov Chain Monte Carlo to simulate the probability measure on one of the groups in the cascade and solve the SDE

$$ds(t) \;=\; -grad[E_{prior}(s) \;+\; E_{likelihood}(s)]dt \;+\; d(W(t)$$

in terms of the d-dimensional Wiener process $W(t)$ and continue iterating until the algorithmic time parameter t is so large that approximate statistical equilibrium has been reached. The previous propagated template, say $I^{\alpha}_{temp}(k,x)$, is then further propagated

$$I^{\alpha}_{temp}(k,x) \rightarrow I^{\alpha}_{temp}(k+1,x) = I^{\alpha}_{temp}(k, s^*_k x)$$

where s^*_k is the resulting group element from the SDE.

We now do this for each group in the cascade, successively propagating the template, see Matejic (1997). The resulting propagated template then induces the output of the vision engine under the adopted strategy for seeing.

3. Experiments.

Based on the above principles three strategies for seeing have been developed. Due to space limitations it is not possible to describe them in detail here; the reader is referred to Grenander (1998) where the strategies are fully described and their code is attached. The algorithms are so complex that it is difficult to predict their behavior. For this reason extensive experimentation has been carried out in order to find their strengths and weaknesses.

Here a few remarks will have to suffice. The first strategy was considering objects as sets in the plane and the attention funcion was then just measuring the optical activity in sub-sets. The observations were degraded by deformations of the generators, additive noise as well as clutter. Additive noise was easily handled by this algorithm, while clutter confused the algorithm and occassionally led it to make the wrong decision; this occurred even for moderate amounts of clutter. Obscuration was well handled if the overlap of generators was not too large but otherwise mistakes were made sometimes.

To handle obscuration better a second strategy was developed where the generators were closed simple curves in the plane, the boundaries of the sets. The attention function was designed to measure the (estimated) lengths of boundaries in subsets. Again additive noise caused no problem for the recognition algorithm. This strategy was less confused by obscuration than the first one, but it was quite sensitive to clutter, apparently because of the differential-geometric nature of the attention function.

A third strategy was constructed for a dynamic situation with moving generators. The attention function measured the amount of change in sub-sets from one frame to the next. This strategy was not very sensitive to clutter although it sometimes made the algorithm answer "do not understand the scene".

We draw the following conclusions from the experiments. The experiments have been carried out under controlled laboratory conditions, and since the inference algorithms are optimal modulo given assumptions, the observed weaknesses of the engines cannot be blamed on the construction of the algorithms. Instead they are essential to the visual set up and point to the need for a careful formulation of the purpose to be realized.
(i) Additive noise in not much of a problem but clutter is. In order to build effective strategies in the future one should *develop a better understanding of how clutter can represented mathematically.*
(ii) The purpose of a vision engine must be clearly articulated with *attention functions that combine several properties of the image*, not just a single one as in the three experiments.
(iii) Related to (ii) is the *need for incorporating cues in the observed "image"* $I^{\mathcal{D}}$: in addition to the image itself relevant facts known to the operator of the inference engine should also be included.
(iv) The vision engines should be *integrated systems* for multi-sensor, multi-target, multi-purpose situations with parallel implementations.

Work is under way to implement (i) - (iv).

References.

G.E. Christensen, R.D. Rabbitt, M.I. Miller (1993): A Deformable Neuroanatomy Textbook Based on Viscous Fluid Mechanics, in Proc. 27th Annual Conf. Information Sci. and Systems, J.Prince and T. Runolfsson eds., pp, 211-216.

J.J. Gibson (1979): The Ecological Approach to Visual Perception, Houghton Mifflin.

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

U.Grenander (1998): Strategies for Seeing, Brown University Tech. Rep.

L. Matejic (1998): Group Cascades for Representing Biological Variability in Medical Images, to appear in the Quart. Appl. Math.

D.Mumford (1994): Neuronal Architectures for Pattern-Theoretic Problems, in Large-Scale Neuronal Theories of the Brain, C.Koch and J.Davis eds., MIT Press.

D.Mumford(1996): The Statistical Description of Visual Signals, in ICIAM95 , K.Kirchgassner, O. Mahrenholtz, R. Mennicken eds., Akademie Verlag.

Ulf Grenander
Brown University Box F
Providence RI 02912
USA