

Pattern Theory: A Unifying Perspective

David Mumford *

1. Introduction

The term “Pattern Theory” was introduced by Ulf Grenander in the 70s as a name for a field of applied mathematics which gave a theoretical setting for a large number of related ideas, techniques and results from fields such as computer vision, speech recognition, image and acoustic signal processing, pattern recognition and its statistical side, neural nets and parts of artificial intelligence (see [Grenander 76–81]). When I first began to study computer vision about ten years ago, I read parts of this book but did not really understand his insight. However, as I worked in the field, every time I felt I saw what was going on in a broader perspective or saw some theme which seemed to pull together the field as a whole, it turned out that this theme was part of what Grenander called pattern theory. It seems to me now that this is the right framework for these areas, and, as these fields have been growing explosively, the time is ripe for making an attempt to reexamine recent progress and try to make the ideas behind this unification better known. This article presents pattern theory from my point of view, which may be somewhat narrower than Grenander’s, updated with recent examples involving interesting new mathematics. I want to define pattern theory as:

the analysis of the patterns generated by the world in any modality, with all their naturally occurring complexity and ambiguity, with the goal of reconstructing the processes, objects and events that produced them.

Thus *vision* usually refers to the analysis of patterns detected in the electromagnetic signals of wavelengths 400-700 nm. incident at a point in space from different directions. *Hearing* refers to the analysis of the patterns present in the oscillations of 60-20,000 hertz in air pressure at a point in space as a function of time, both with and without human language. We may also say that *medical expert systems* are concerned with the analysis of the patterns in the symptoms, history and tests presented by a patient: this is a higher level modality, but still one in which the world

* Supported in part by NSF Grant DMS 91-21266 and by the Geometry Center, University of Minnesota, a STC funded by NSF, DOE and Minnesota Technology Inc.

generates confusing but structured data from which a doctor seeks to infer hidden processes and events. *Touch*, especially in conjunction with active motor control, either in an animal or robot, is yet another such channel.

Let me give two examples to help fix ideas. Figure 1 shows the graph of the pressure $p(t)$ while the word “SKI” is being pronounced. Note how the signal shows four distinct wave forms: something close to white noise during the pronunciation of the sibilant “S”, then silence followed by a burst which conveys the plosive “K”, then an extended nearly musical note for the vowel “I”. The latter has a fundamental frequency corresponding to the vibration of the vocal cords, with many harmonics whose power peaks around three higher frequencies, the formants. Finally, the amplitude of the whole is modulated during the pronunciation of the word.

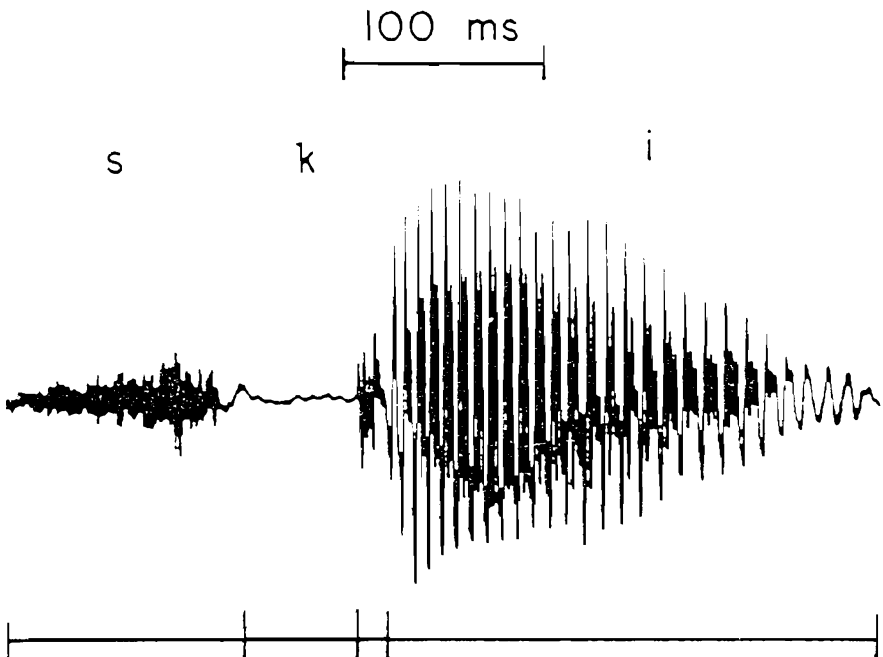


Figure 1. Acoustic waveform for an utterance of the word *SKI*

In this example, the goal of perceptual signal processing is to identify these four wave forms, characterize each in terms of its frequency power spectrum, its frequency and amplitude modulation, and then, drawing on a memory of speech sounds, identify each wave form as being produced by

the corresponding configurations of the speaker's vocal tract, and finally, label each with its identity as an English phoneme. In addition, one would like to describe explicitly the stress, pitch and quality of the speaker's voice, using this later to help disambiguate the identity of the speaker and the intent of the utterance.

Figure 2a shows the graph of the light intensity $I(x, y)$ of a picture of a human eye: it would be hard to recognize this as an eye, but the black and white image defined by the same function is shown in Figure 2b.

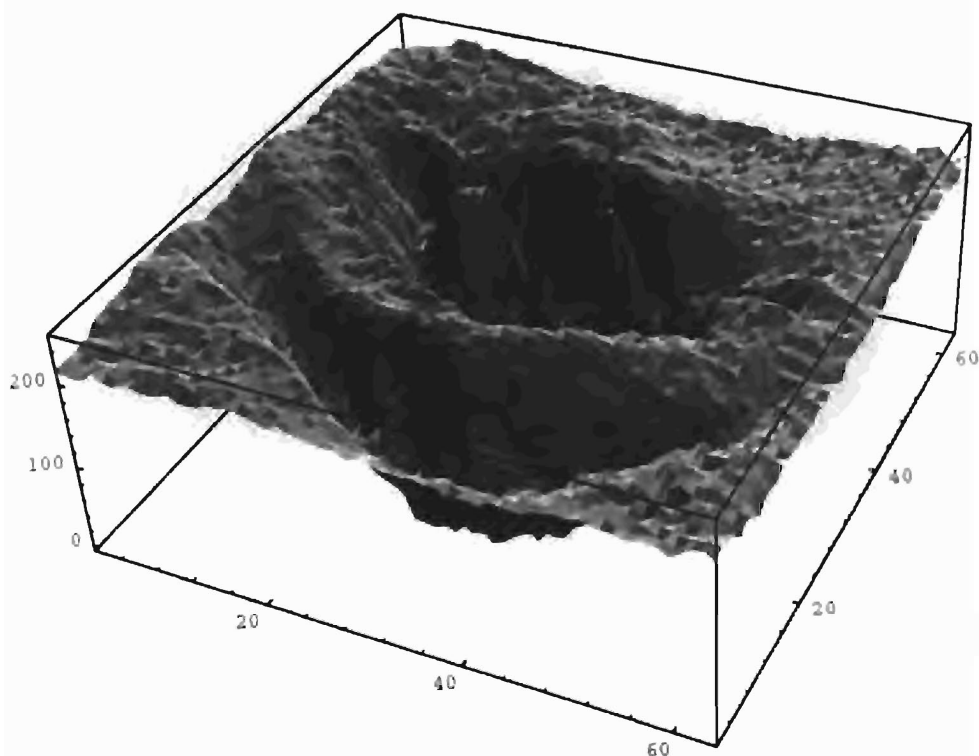


Figure 2a. Visual waveform for an image of an eye

Note again how the domain of the signal is naturally decomposed into regions where I has different values or different spatial frequency behavior: the pupil, the iris, the whites of the eyes, the lashes, eyebrows and skin. The goal of perceptual signal processing is again to describe this function of two variables as being built up from simpler signals on subdomains on which it either varies slowly or is statistically regular, i.e., approximately stationary. These statistics may be its spatial power spectrum or a method of generation from elementary units called *textons* by repetition with various

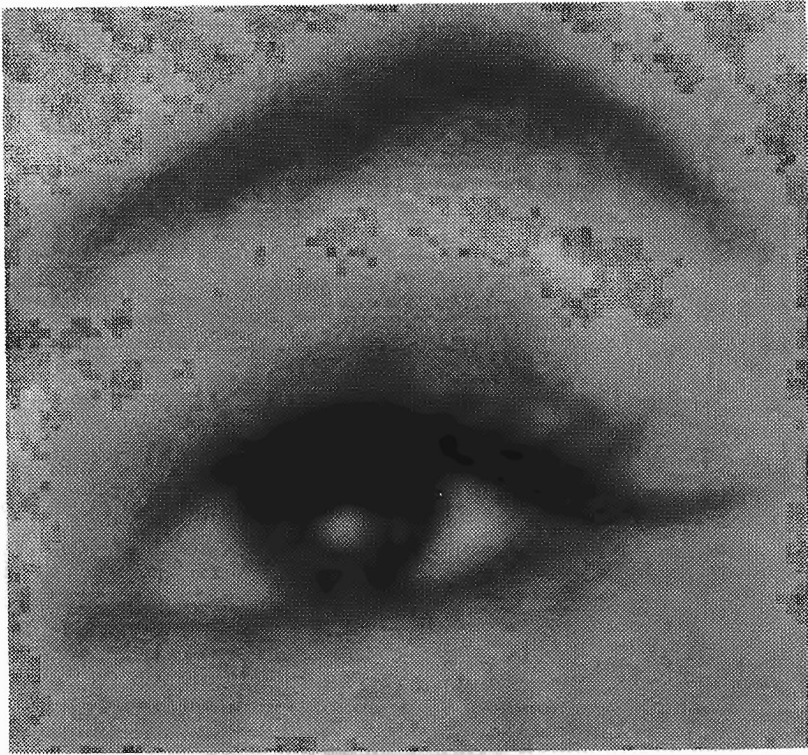


Figure 2b. Identical waveform presented using variable intensity

modifications. These modifications in particular include spatial distortion, contrast modulation and interaction with larger scale structures. This description of the signal may be computed either prior to or concurrent with a comparison of the signal with remembered eye shapes, which include a description of the expected range of variation of eyes, specific descriptions of the eyes of well-known people, and so on.

In order to understand what the field of pattern theory is all about, it is necessary to begin by addressing a major misconception, namely, that the whole problem is essentially trivial. The history of computer speech and image recognition projects, like the history of AI, is a long one of ambitious projects which attained their goals with carefully tailored artificial input but which failed as soon as more of the complexities of real world data were present. The source of this misconception, I believe, is our subjective impression of perceiving instantaneously and effortlessly the significance of the patterns in a signal, e.g. the word being spoken or which face is being seen.

Many psychological experiments, however, have shown that what we

perceive is not the true sensory signal, but a rational reconstruction of what the signal should be. This means that the messy ambiguous raw signal never makes it to our consciousness but gets overlaid with a clearly and precisely patterned version whose computation demands extensive use of memories, expectations and logic. An example of how misleading our impressions are is shown in Figure 3 below.



Figure 3. A challenging image for computers to recognize

The reader will instantly recognize that it is an image of an old man sitting on a park bench. But ask yourself — how did you know that? His face is almost totally obscure, with his hand merging with his nose: the most distinct shape is that of his hat, which by itself could be almost anything; even his jacket merges in many places with the background because of its

creasing and the way light strikes them, so no simple algorithm is going to trace its contour without getting lost. However, when you glance at the picture, in your mind's eye, you "see" the face and its parts distinctly; the man's jacket is a perfectly clear coherent shape whose creases, instead of confusing you, in fact contribute to your perception of its 3-dimensional structure. The ambiguities which must have been present in the early stages of your processing of this image never become conscious because you have found an explanation of every peculiarity of the image, a match with remembered shapes and lighting effects. In fact, the problems of pattern theory are hard, and although major progress has been made in both speech and vision in the last 5 years, a robust solution has not been achieved!

2. Mathematical formulation of the field

To make the field of pattern theory precise, we need to formulate it mathematically. There are three parts to this which were all made quite clear by Grenander: the first is the description of the players in the field, the fundamental mathematical objects which will appear in each case. The second is to restrict the possible generality of these objects by using something about the nature of the world. This gives us a more circumscribed, more focussed set of problems to study. Finally, since the goal of the field is the reconstruction of hidden facts about the world, we aim primarily for algorithms, not theorems; and the last part is the general framework for these algorithms. In this section, we look at the first part, the basic mathematical objects of pattern theory.

For this there are two essentially equivalent formulations, one using Bayesian statistics and one using information theory. The statistical approach (see for instance [D. Geman 91]) is this: consider all possible signals $f(\vec{x})$ which may be perceived. These may be considered as elements of a space Ω_{obs} of functions $f : D \rightarrow V$. For instance, speech is defined by pressure $p : [t_1, t_2] \rightarrow \mathbf{R}_+$ as a function of time, color vision is defined by intensity I on a domain $D \subset S^2$ of visible rays with values in the convex cone of colors $V_{\text{RGB}} \subset \mathbf{R}^3$, or these functions may be sampled on a discrete subset of the above domains, or their values may be approximated to finite precision, etc.

The first basic assumption of the statistical approach is that nature determines a probability p_{obs} on a suitable σ -algebra of subsets of Ω_{obs} , and that, in life, one observes random samples from this distribution. These signals, however, are highly structured as a result of their being produced by a world containing many processes, objects and events which do not

appear explicitly in the signal. This means many more random variables are needed to describe the state of the world. The second assumption is that the possible states w of the world form a second probability space Ω_{wld} and that there is a big probability distribution $p_{o,w}$ on $\Omega_{\text{obs}} \times \Omega_{\text{wld}}$ describing the probability of both observing f and the world being in state w . Then p_{obs} should be the marginal distribution of $p_{o,w}$ on Ω_{obs} . The goal of pattern theory is to infer the state of the world w , given the measurement f , and for this we may use Bayes's rule:

$$p(w|f) = \frac{p(f|w) \cdot p(w)}{p(f)} \tag{1}$$

leading to the *maximum likelihood reconstruction of the state of the world*:

$$\text{ML estimate of } w = \arg \max_w [p(f|w) \cdot p(w)] \tag{2}$$

The statistical approach entails constructing the probability space $(\Omega_{\text{obs}} \times \Omega_{\text{wld}}, p_{o,w})$ and finding algorithms to compute the ML-estimate.

In the information theoretic approach (see for instance [Rissanen 89]) we assume D and V , hence Ω_{obs} are finite. The idea is that instead of writing out any particular perceptual signal f in raw form as a table of values, we seek a method of encoding f which minimizes its expected length in bits: i.e., we take advantage of the patterns possessed by most f to encode them in a compressed form. We consider coding schemes which involve choosing various auxiliary variables w and then encoding the particular f using these w (e.g., w might determine a specific typical signal f_w and we then need only to encode the deviation $(f - f_w)$). We write this:

$$\text{length}(\text{code}(f, w)) = \text{length}(\text{code}(w)) + \text{length}(\text{code}(f \text{ using } w)). \tag{3}$$

It might appear that such optimal encodings of signals would lead you to odd combinatorial schemes that have nothing to do with what is actually happening in the world. Remarkably, this isn't the case and, in fact, it seems to lead you *automatically, without prior knowledge of the world*, to the same hidden variables on which the Bayesian theory is based. This link between the two approaches comes from Shannon's optimal coding theorem. This theorem states that, given a class of signals f , the coding scheme for such signals for which a random signal has the smallest expected length satisfies

$$\text{len}(\text{code}(f)) = -\log_2 p(f) \tag{4}$$

(where fractional bit lengths are achieved by actually coding several f 's at once, and in doing this, the LHS gets asymptotically close to the RHS when longer and longer sequences of signals are encoded at once). Using Shannon's theorem, and taking \log_2 of (2), we get the *minimum description length reconstruction of the world*:

$$\text{MDL est. of } w = \arg \min_w [\text{len}(\text{code}(w)) + \text{len}(\text{code}(f \text{ using } w))]. \quad (5)$$

The information-theoretic approach entails constructing a coding scheme $\{w\}$ and finding algorithms to compute (5). Its great strength is that, as opposed to the Bayesian approach, it does not require a prior knowledge of the physics, chemistry, biology, sociology, etc. of the world, but gives you a way of discovering these facts. In Section 5.4, we will give an example of how this works.

3. Four universal transformation of perceptual signals

The above formulation of pattern theory provides a framework in which to analyze signals, but it says nothing about the nature of the patterns which are to be expected, what distortions, complexities and ambiguities are to be expected, hence what kinds of probability spaces Ω_{obs} are we likely to encounter, how shall we encode them, etc.

What gives the field its characteristic flavor is that the world does not have an infinite repertoire of different tricks which it uses to disguise what is going on. Consider the coding schemes used by engineers for the transmission of electrical signals. They use a small number of well-defined transformations such as AM and FM encoding, pulse coding, etc. to convert information into a signal which can be efficiently communicated. Analogous to this, the world produces sound to be heard, light to be seen, surfaces to be felt, and so on, which are all, in various ways, reflections of its structure.

We may think of these signals as the productions of a particularly perverse engineer, who presents us with the problem of decoding this message, e.g., of recognizing a friend's face or estimating the trajectory of oncoming traffic, etc. The second contention of pattern theory is that such signals are derived from the world by *four types of transformations or deformations*, which occur again and again in different guises. The bad news is that these four transformations produce much more complex effects than the coding schemes of engineers, hence the difficulty of decoding them by the standard tricks of electrical engineering. The good news is that these transformations are not arbitrary recursive operations which produce unlearnable complexity. For instance, in the formal study of language learnability,

Gold's theorem gives very strong restrictions on what can be learned (see [Osherson-Weinstein 84] for an excellent exposition). But the study of perceptual signals suggests that this is largely irrelevant, that the languages in which perceptual signals speak are of very special types. In the terminology of [Grenander 76], simple unambiguous signals from the world are referred to as *pure images* and the transformations on them are called *deformations*, which produce the actually observed perceptual signals which he called *deformed images*.

The four transformations that I propose as the basic types occurring in natural perceptual signals are:

1. *Noise and blur*. These effects are the bread and butter of standard signal processing, caused for instance by sampling error, background noise and imperfections in your measuring instrument such as imperfect lenses, veins in front of the retina, dust and rust. A typical form of this transformation is given by

$$I \rightarrow (I * \sigma)(x_i) + n_i \quad (6)$$

where σ is a blurring kernel, x_i are the points where the signal is sampled and n_i is some kind of additive noise, e.g., Gaussian, but of course much more complex formulac are possible. Especially significant is that Gaussian noise is usually a poor model of the noise effects, for example when the noise is caused by finer detail which is not being resolved. Rosenfeld calls such an *n clutter*, which conveys what it often represents. These transformations are part of what Grenander calls *changes in contrast*. When they are present, the unblurred noiseless I should be one of the variables w as getting rid of noise and blur reveals the hidden processes of the world more clearly.

2. *Multi-scale superposition*. Signals typically reveal one set of structures caused by one set of phenomena in the world when analyzed locally, at high precision, and other structures and phenomena when analyzed globally and coarsely, at low precision. For instance in images, local properties include sharp edges, texture details and local irregularities of shapes, which coexist with global properties like slowly varying shading or texture statistic gradients and the overall shape of an object. In speech, information is conveyed by the highest frequency formants, by the lower frequency vibration of the vocal cords and the even slower modulation of stress. A typical form of this transformation is given by

$$I_1, I_2 \longrightarrow (I_1 + I_2) \quad \text{or} \quad (I_1 \cdot I_2) \quad \text{or} \quad \sigma(I_1, I_2) \quad (7)$$

where I_1 and I_2 represent band pass signals in disjoint frequency

bands, which can be combined either additively (the usual superposition of high and low frequency effects), multiplicatively (as in amplitude modulation of a carrier signal for example) or by some more complex rule σ . This type of deformation does not seem to have been made explicit by Grenander. The individual components I_k of I should be included in the variables w .

3. *Domain warping.* Two signals generated by the same object or event in different contexts typically differ because of expansions or contractions of their domains, possibly at varying rates: phonemes may be pronounced faster or slower, the image of a face is distorted by varying expression and viewing angle. In speech, this is called “time warping” and in vision, this is modeled by “flexible templates”. In both cases, a diffeomorphism of the domain of the signal brings the variants much closer to each other so that this transformation is given by

$$I \longrightarrow (I \circ \psi) \tag{8}$$

where ψ represents a diffeomorphism of the domain of I . These transformations are what Grenander calls *background deformations*. The diffeomorphism ψ should be one of the variables w .

4. *Interruptions.* Natural signals are usually analyzed best after being broken up into pieces consisting of their restrictions to subdomains. This is because the world itself is made up of many objects and events and different parts of the signal are caused by different objects or events. For instance, an image may show different objects partially occluding each other at their edges, as in Figure 3 where the old man is an object which occludes part of the park bench. In speech, the phonemes naturally break up the signal and, on a larger scale, one speaker or unexpected sound may interrupt another. Such a transformation is given by such a formula as

$$I_1, I_2 \longrightarrow (I_1|_{D'}, I_2|_{D-D'}) \tag{9}$$

where I_2 represents the background signal which is interrupted by the signal I_1 on a part D' of its domain D , (or I_2 may only be defined on $D - D'$). This type of deformation is called *incomplete observations* by Grenander. The components I_k and the domain D' should be included in the variables w .

What makes pattern theory hard is not that any of the above transformations are that hard to detect and decode in isolation, but rather that all four of them tend to coexist, and then the decoding becomes hard.

4. Pattern analysis cannot be done without pattern synthesis

Taking the Bayesian definition of the objects of pattern theory, we note that the probability distribution $(\Omega_{\text{obs}} \times \Omega_{\text{wld}}, p_{o,w})$ allows us to do two things. On the one hand, we can use it to define the ML-estimate of the state of the world; but we can also sample from it, possibly fixing some of the world variables w , using this distribution to construct sample signals f generated by various classes of objects or events. A good test of whether your prior has captured all the patterns in some class of signals is to see if these samples are good imitations of life. For this reason, Grenander's idea was that the analysis of the patterns in a signal and the synthesis of these signals are inseparable problems: computer vision should not be separated from computer graphics, nor speech recognition from speech generation. This is the third part of our definition of pattern theory. What gives it force is the idea of constraining not merely your theory but also your *algorithms* to require that they explicitly model the universal transformations, hence can be used to generate signals as well as analyze them.

Many of the early algorithms in pattern recognition were purely *bottom-up*. For example, one class of algorithms started with a signal, computed a vector of "features", numerical quantities thought to be the essential attributes of the signal, and then compared these feature vectors with those expected for signals in various categories. This was used to classify images of alpha-numeric characters or phonemes, for instance. Such algorithms give no way of reversing the process, of generating typical signals. The problem these algorithms encountered was that they had no way of dealing with anything unexpected, such as a smudge on the paper obscuring a character, or a cough in the middle of speech. These algorithms did not say what signals were expected, only what distinguished typical signals in each category.

In contrast, a second class of algorithms works by actively reconstructing the signal being analyzed. In addition to the bottom-up stage, there is a *top-down* stage in which a signal with the detected properties is synthesized and compared to the present input signal. What needs to be checked is whether the input signal agrees with the synthesized signal to within normal tolerances, or whether the residual is so great that the input has not been correctly or fully analyzed. This architecture is especially important for dealing with the fourth type of transformation: interruptions. When these are present, the features of the two parts of the signal get confused. Only when the obscuring signal is explicitly labelled and removed, can the features of the background signal be computed. We may describe this top-down stage as "pattern reconstruction" in distinction to the bottom-up

purely pattern recognition stage. A flow chart for such algorithms is shown in Figure 4.

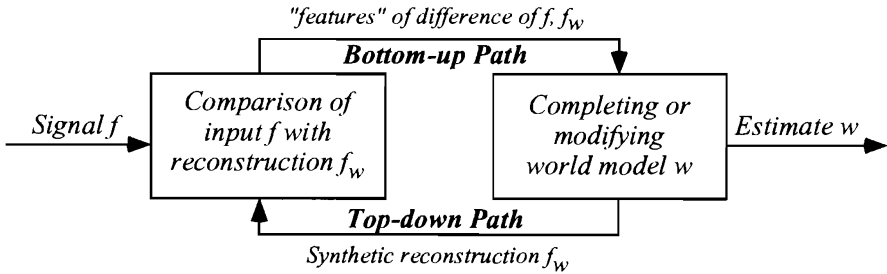


Figure 4. The fundamental architecture of Pattern Theory

We will not develop this aspect of pattern theory in this paper, but would like to mention briefly several papers where these ideas are developed. A strong argument for the necessity of a top-down stage for the recognition of heavily degraded signals, such as faces in deep shadow, is given in [Cavanagh 91]. Neural net theory has gone in several directions: while “feed-forward” nets categorize in an exclusively bottom-up manner, the “attractive neural nets” with symmetric connections ([Hopfield 82], [D. Amit 89]) seek not merely to categorize but also to construct the prototype ideal version of the category by a kind of pattern completion which they call “associative memory”. What these nets do not do is to go back and attempt to compare this reconstruction with the actual input to see if the full input has been “explained”. One demonstration system that does this is Grossberg and Carpenter’s “adaptive resonance theory” ([Carpenter-Grossberg 87]). A proposal for the neuroanatomical substrate for such bottom-up/top-down loops in mammalian cortex is put forth in [Mumford 91-92].

5. Examples

In this section, we want to present several examples of interesting mathematics which have come out of pattern theory, in attempting to come to grips with one or another of the above universal transformations. These examples are from vision because this is the field I know best, but many of these ideas are used in speech recognition too.

5.1 Pyramids and wavelets

The problem of detecting transformations of the second kind, i.e., of analyzing functions that convey information on more than one scale, has arisen in many disciplines. The classical method of separating additively combined

scales is, of course, Fourier analysis. But what is usually required is to analyze a function locally both in its original domain *and* in the domain of its Fourier transform, and Fourier analysis does not do this. In computer vision, at least as far back as the early 70s, this problem led to the idea of analyzing an image by means of a “pyramid”, e.g., [Uhr 72], [Rosenfeld-Thurston 71]. In its original incarnation, the main idea was to compute a series of progressively coarser resolution images by blurring and resampling, e.g., a set of $(2^n \times 2^n)$ -pixel images, for $n = 10, 9, \dots, 1$. Putting these together, the resulting data structure looks like an exponentially tapering pyramid. Instead of running algorithms that took time proportional to the width of the image, one ran the algorithms up and down the pyramid, possibly in parallel at different pixels, in time proportional to $\log(\text{width})$. Typical algorithms that were studied at this time are morphological ones, involving for instance linking and marking extended contours, which have nothing to do with filtering or linear expansions. The bottom layer of this so-called *Gaussian pyramid* held the original image, with both high and low frequency components, although it was used only to add local or high-frequency information.

In the early 80s, the idea of using the pyramid to separate band pass components of a signal and thus to expand that signal arose both in computer vision [Burt-Adelson 83] (where they *subtracted* successive layers of the Gaussian pyramid, producing what they called the *Laplacian pyramid*) and in petroleum geology [Grossman-Morlet 84]. Figure 5 shows this Laplacian pyramid for a face image: note that the high-frequency differences show textures and sharp edges, while the low frequency differences show large shapes.

This work led directly to the idea of wavelets and wavelet expansions which now seem to be the most natural way to analyze a signal locally in both space and frequency. Mathematically, the idea is simply to expand an arbitrary function $f(\vec{x})$ of n variables as a sum:

$$f(\vec{x}) = \sum_{\text{scale } k \in Z} \left[\sum_{\vec{n} \in \text{lattice } L} \sum_{\text{fin. \# of } \alpha} a_{k, \vec{n}, \alpha} \psi_{\alpha}(\lambda^k \vec{x} + \vec{n}) \right] \quad (10)$$

where the ψ_{α} are suitable functions, called wavelets, with mean 0. Usually $\lambda = 2$, and, at least in dimension 1, there is a single α and wavelet ψ_{α} . The original expansions of Burt and Adelson, which are not quite of this form, have been reinvestigated from a more mathematical point of view in [Mallat 89]. The basic link between the expansion in (10) and pyramids is this: define a space V_m to be the set of f 's whose expansions involve only terms with $k \leq m$. This defines a “multi-resolution ladder” of subspaces

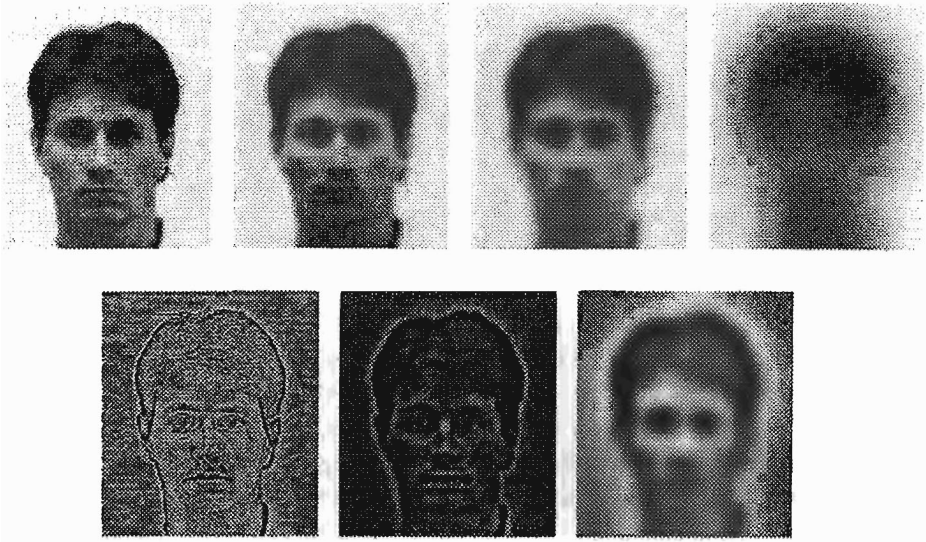


Figure 5. The Gaussian and Laplacian pyramids for a face image of functions with more and more detail:

$$\dots \subset V_{-1} \subset V_0 \subset V_1 \subset \dots \subset L^2(\mathbf{R}^n) \quad (11)$$

such that $f(x) \mapsto f(2x)$ maps V_m isomorphically onto V_{m+1} . Then one may think of V_m as functions which have been blurred and sampled at a spacing 2^{-m} : i.e., the level of the pyramid of $(2^m \times 2^m)$ -pixel images. The mathematical development of the theory of these expansions is due especially to Meyer and Daubechies (see [Meyer 86], [Daubechies 88], [Daubechies 90]), who showed that (i) with *very* careful choice of ψ , this expansion is even an orthogonal one, (ii) for many more ψ , the functions on the right form an unconditional but not orthogonal basis of $L^2(\mathbf{R}^n)$ and (iii) for even more ψ , the functions on the right form a “frame”, a set of functions that spans $L^2(\mathbf{R}^n)$ and gives a canonical though non-unique expansion of every f .

From the perspective of pattern theory, we want to make two comments on the theory of wavelets. The first is that they fit in very naturally with the idea of minimum description length. Looked at from the point of view of optimal linear encoding of visual and speech signals (i.e., encoding by linear combinations of the function values), the idea of wavelet expansions is very appealing. This was pointed out early on by Burt and Adelson and data compression has been one of the main applications of wavelet theory ever since. Moreover, its further development leads beyond the classical idea of expanding a function in terms of a fixed basis to the idea

of using a much larger spanning set which *oversamples* a function space and using suitably chosen subsets of this set in terms of which to expand or approximate the given function (see [Coifman-Meyer-Wickerhouser 90] where *wavelet libraries* are introduced). Even though the data needed to describe this expansion or approximation is now both the particular subset chosen and the coefficients, this may be a more efficient code. If so, it should lead us to the correct variables w for describing the world (cf. Section 2): for example, expanding a speech signal using wavelet libraries, different bases would naturally be used in the time domains during which different phonemes were being pronounced – thus the break-up of the signal into phonemes is discovered as a consequence of the search for efficient coding! It also appears that nature uses wavelet type encoding: there are severe size restrictions on the optic nerve connecting the retina with the higher parts of the brain and the visual signal is indeed transmitted using something like a Burt-Adelson wavelet expansion [Dowling 87].

The second point is that wavelets, even in their oversampled form, are still just the linear side of pyramid multi-scale analysis. In our description of multi-scale transformations of signals in Section 3, we pointed out that the two scales can be combined by multiplication or a more general function σ as well as by addition. To decode such a transformation, we need to perform some local non-linear step, such as rectification or auto-correlation, at each level of the pyramid before blurring and resampling. An even more challenging and non-linear extension is to a *multi-scale description of shapes*: e.g., subsets $S \subset \mathbf{R}^2$ with smooth boundary. The analog of blurring a signal is to let the boundary of S evolve by diffusion proportional to its curvature (see [Gage-Hamilton 86], [Grayson 87]). Although there is no theory of this at present, one should certainly have a multi-scale description of S starting from its coarse diffused form – which is nearly round – and adding detailed features at each scale. In yet another direction, face recognition algorithms have been based on matching a crude blurry face template at a low resolution, and then refining this match, especially at key parts of the face like the eyes. This is the kind of general pyramid algorithm that Rosenfeld proposed many years ago, many of which have been successfully implemented by Peter Burt and his group at the SRI Sarnoff Laboratory.

5.2 Segmentation as a free-boundary value problem

A quite different mathematical theory has arisen out of the search for algorithms to detect transformations of the fourth kind, interruptions. Evidence for an interruption or a discontinuity in a perceptual signal comes from two sources: the relative homogeneity of the signal on either side of the boundary and the presence of a large change in the signal across the

boundary. One approach is to model this as a variational problem: assuming that a blurred and noisy signal f is defined on a domain $D \subset \mathbf{R}^n$, one seeks a set $\Gamma \subset D$ and a smoothed version g of f which is allowed to be discontinuous on Γ such that:

- g is as close as possible to f ,
- g has the smallest possible gradient on $D - \Gamma$,
- Γ has the smallest possible $(n - 1)$ -volume.

These conditions define a variational problem, namely to minimize the functional

$$E(g, \Gamma) = \mu^2 \int_D \dots \int (f - g)^2 + \int_{D - \Gamma} \dots \int \|\nabla g\|^2 + \nu |\Gamma| \quad (12)$$

where μ and ν are suitable constants weighting the three terms and $|\Gamma|$ is the $(n - 1)$ -volume of Γ . The g minimizing E may be understood as the optimal piecewise smooth approximation to the quite general function f . In Grenander's terms, the function g is the pure image and f is the deformed image; I like to call g a *cartoon* for the signal f . The Γ minimizing E is a candidate for the boundaries of parts of the domain D of f where different objects or events are detected. Segmenting the domain of perceptual signals by such variational problems was proposed independently by S. and D. Geman and by A. Blake and A. Zisserman (see [Geman-Geman 84] and [Blake-Zisserman 87]) for functions on discrete lattices, and was extended by [Mumford-Shah 89] to the continuous case.

In the case of visual signals, the domain D is 2-dimensional and we want to decompose D into the parts on which different objects in the world are projected. When you reach the edge of an object as seen from the image plane, the signal f typically will be more or less discontinuous (depending on noise and blur and the lighting effects caused by the grazing rays emitted by the surface as it curves away from the viewer). An example of the solution of this variational problem is shown in Figure 6: Figure 6a is the original image of the eye, 6b shows cartoon g and 6c shows the boundaries Γ . This is a case where the algorithm succeeds in finding the "correct" segmentation, but it doesn't always work so well.

Figure 7 gives the same treatment as Figure 6, to the "oldman" image. Note that the algorithm fails to find the perceptually correct segmentation in several ways: the man's face is connected to his black coat and the black bar of the bench and the highlights on the back of his coat are treated as separate objects. One reason is that the surfaces of objects are often textured, hence the signal they emit is only statistically homogeneous. More sophisticated variational problems are needed to segment textured objects (see below).



Figure 6. Segmentation of the eye-image via optimal piecewise smooth approximation



Figure 7. Segmentation of the oldman-image via optimal piecewise smooth approximation

From a mathematical standpoint, it is important to know if this variational problem is well-posed. It has been proven that E has a minimum if Γ is allowed to be a closed rectifiable set of finite Hausdorff $(n - 1)$ -dimensional measure and g is taken in a certain space SBV (“special bounded variation”, which means that the distributional derivative of g is the sum of an L^2 -vector field plus a totally singular distribution supported on Γ) (see [DeGiorgi-Carricco-Leaci 88], [Ambrosio-Tortorelli 89] and [Dal Maso-Morel-Solimini 89]). Unfortunately, it seems hard to check whether these minima are “nice” when f is, e.g., whether, when $n = 2$, Γ is made up of a finite number of differentiable arcs, though Shah and I have conjectured that this is true. Of course, if the signal is replaced by a sampled version, the problem is finite dimensional and certainly well-posed.

This variational problem fits very nicely into both the Bayesian framework and the information theoretic one. Geman and Geman introduced it for discrete domains in the Bayesian setting. The basic idea is to define probability spaces by Gibbs fields. Let $D = \{x_\alpha\}$ be the domain, $\{f_\alpha\}$ and $\{g_\alpha\}$ the values of f and g at x_α . To describe Γ , for each pair of “adjacent pixels” α and β , let $\ell_{\alpha,\beta} = 1$ or 0 depending on whether or not Γ separates the pixels α and β : these random variables are called the *line process*. Then we define a prior probability distribution on the random variables $\{\ell_{\alpha,\beta}\}$ by the formula

$$p(\{\ell_{\alpha,\beta}\}) = \frac{e^{-\nu(\sum \ell_{\alpha,\beta})}}{Z_1} \tag{13}$$

where Z_1 is the usual normalizing constant. This just means that boundaries Γ get less and less probable, the bigger they are. Next, we put a conditional probability distribution on $\{g_\alpha\}$ conditional on the line process by the formula

$$p(\{g_\alpha\}|\{\ell_{\alpha,\beta}\}) = \frac{e^{-\sum_{\alpha,\beta \text{ adj}} (1-\ell_{\alpha,\beta}) \cdot (g_\alpha - g_\beta)^2}}{Z_2}. \tag{14}$$

This is a discrete form of the previous E : if adjacent pixels α and β are *not* separated by Γ , then $\ell_{\alpha,\beta} = 0$ and the probability of $\{g_\alpha\}$ goes down as $|g_\alpha - g_\beta|$ gets larger, but if they *are* separated, then $\ell_{\alpha,\beta} = 1$ and g_α and g_β are independent. Together, the last two equations define an intuitive prior on $\{g_\alpha, \ell_{\alpha,\beta}\}$ enforcing the idea that g is smooth except across the boundary Γ . The data term in the Bayesian approach makes the observations $\{f_\alpha\}$ equal to the model $\{g_\alpha\}$ plus Gaussian noise, i.e., it defines the conditional

probability by the formula

$$p(\{f_\alpha\}|\{g_\alpha, \ell_{\alpha,\beta}\}) = \frac{e^{-\mu^2 \cdot \sum_\alpha (f_\alpha - g_\alpha)^2}}{Z_3}. \quad (15)$$

Multiplying (12), (13) and (14) defines a probability space ($\Omega_{\text{obs}} \times \Omega_{\text{wid}}, p_{o,w}$) as in section 2 and taking $-\log$ of this probability, we get back a discrete version of E up to a constant. Thus the ML-estimate of the world variables $\{g_\alpha, \ell_{\alpha,\beta}\}$ is essentially the minimum of the functional E .

This probability space is closely analogous to that introduced in physics in the Ising model. In terms of this analogy, the discontinuities Γ of the signal are exactly the phase transitions of statistical mechanics.

From the information-theoretic perspective, we want to interpret E as the bit length of a suitable encoding of the image $\{f_\alpha\}$. These ideas have not been fully developed, but for the simplified model in which Γ is assumed to divide up the domain into pieces $\{D_k\}$ on which the image is approximately a constant $\{g_k\}$, this interpretation was pointed out by [Leclerc 89]. We imagine encoding the image by starting with a ‘‘chain code’’ for Γ : the length of this code will be proportional to its length $|\Gamma|$. Then we encode the constants $\{g_k\}$ up to some accuracy by a constant times the number of these pieces k . Finally, we encode the deviation of the image from these constants by Shannon’s optimal encoding based on the assumption that $f_\alpha = g_k + \text{Gaussian noise } n_\alpha$. The length of this encoding will be a constant times the first term in E . (If g is not locally constant, we may go on to interpret the second term in E as follows: consider the Neumann boundary value problem for the laplacian Δ acting on the domain $D - \Gamma$. We may expand g in terms of its eigenfunctions, and encode g by Shannon’s optimal encoding assuming these coefficients are independently normally distributed with variances going down with the corresponding eigenvalues. The length will be this second term.)

Many variants of this Gibbs field or ‘‘energy functional’’ approach to perceptual signal processing have been investigated. Some of these seek to incorporate texture segmentation, e.g., [Geman-Geman-Graffigne-Dong 90] and [Lee-Mumford-Yuille 92] (which proposes an algorithm that should also segment most phonemes in speech) and others to deal with the asymmetry of boundaries caused by the 3D-world: at a boundary, one side is in front, the other in back [Nitzberg-Mumford 90]. The ‘‘Hidden Markov Models’’ used in speech recognition are Gibbs fields of this type. To clarify the relationship, recall that HMM’s are based on modelling speech by a Markov chain whose underlying graph is made up of subgraphs, one for each phoneme and whose states predict the power spectrum of the speech signal in local time intervals. Assuming a specific speech signal f is being

modelled, HMM-theory computes the ML sample of this chain conditional on the observed power spectra. Note that any sample of the chain defines a segmentation of time by the set $\Gamma = \{t_k\}$ of times at which the sample moves from the subchain for one phoneme to another, and each interval $t_k \leq t \leq t_{k+1}$ is associated to a specific phoneme a_k . Let A be the string $\{a_1 a_2, \dots, a_N\}$. Taking $-\log$ of the probability, the ML estimate of the chain is the pair $\{\Gamma, A\}$ minimizing an energy E of the form

$$E(A, \Gamma) = \sum_k \text{dist.}(f|_{t_k}^{t_{k+1}}, \text{phoneme } a_k) + \nu|\Gamma|, \quad (16)$$

which is clearly analogous to the E 's defined above.

Finally, some physiological theories have been proposed in which various areas of cortex (e.g., V1 and V2) compute the segmentation of images by an algorithm analogous to minimizing (11) [Grossberg-Mignolla 85]. It has also been used in computing depth from stereo [Belhumeur-Mumford 92], [Geiger-Ladendorf-Yuille 92], computing the so-called optical flow field, the vector field of moving objects across the focal plane [Yuille-Grzywacz 89], [Hildreth 84] and many other applications.

We have not mentioned the problem of actually computing or approximating the minimum of energy functionals like E . Four methods have been proposed: in case $n = 1$, we can use *dynamic programming* to find the global minimum fast and efficiently. This applies to the speech applications and is one reason why speech recognition is considerably ahead of image analysis. For any n , [Geman-Geman 84] applied a Monte Carlo algorithm due to [Kirkpatrick-Geloti-Vecchi 83] known as *simulated annealing*. Making this work is something of a black art, as the theoretical bounds on its correctness are astronomical; still it is always an easy thing to try as a first step.

A third method, which seems the most reliable at this point, is the *graduated non-convexity* method introduced in [Blake-Zisserman 87]. It is based on putting the functional E in a family E_t such that $E = E_0$ and E_1 is a convex functional, hence has a unique local minimum. One then starts with the minimum of E_1 and follows it as $t \rightarrow 0$. The final idea is related to the third and that is to use *mean field theory* as in statistical physics: this often leads to approximations to the Gibbs field which allow us to put E in a family becoming convex in the limit (see [Geiger-Yuille 89]).

5.3 Random diffeomorphisms and template matching

The third example concerns the identification of objects in an image, putting them in categories such as “the letter A”, “a hammer” or “my Grandmother’s face”. One of the biggest obstacles in these problems is the

variability of the shapes which belong to such categories. This variability is caused, for example, by changes in the orientation of the object and the viewpoint of the camera, changes in individual objects such as varying expressions on a face and differences between objects of the same category such as different fonts for characters, different brands of hammer, etc. If the shapes were not too variable, one could hope to introduce average examples of each letter, of each tool, of the faces of everyone you know — “templates” for each of these objects — and recognize each such object as it is perceived by comparing it to the various templates stored in memory. Unfortunately, the variations are usually too large for this to work, and, worse than that, some variations occur commonly, while others do not (e.g., faces get wrinkled but never become wavy like water). What we need to do is to explicitly model the common variations and use our knowledge to see if a suitably varied template fits! A large part of this variation can be modelled by domain warping, the third of the transformations introduced in Section 3 and this leads to the study of *deformable templates*, templates whose parts can be changed in size and orientation and shifted relative to each other. These were first introduced in computer vision by [Fischler-Elschlager 73] who called them “templates with springs” but the idea is well-known in biology, e.g., in the famous and beautiful book [Thompson 17] (see Figure 8a, showing the deformations between three primate skulls).

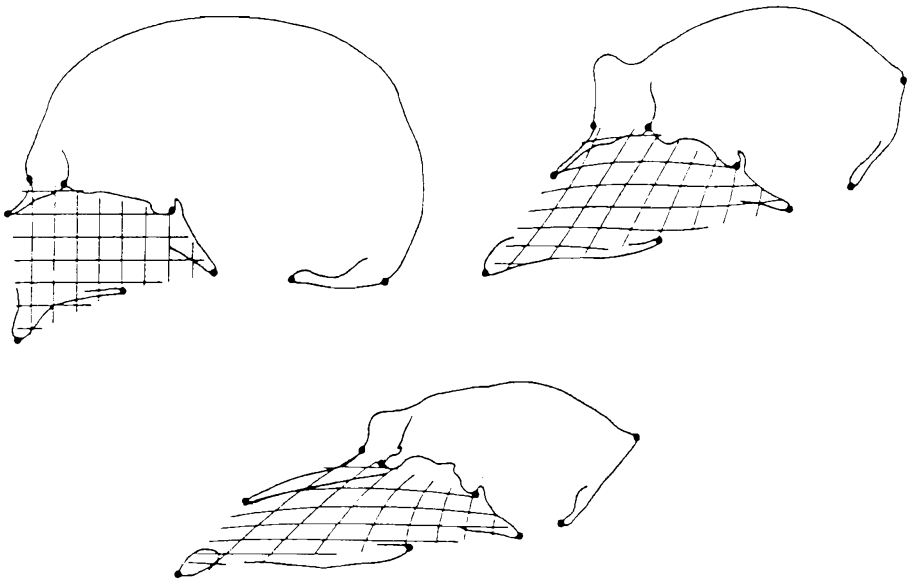


Figure 8a. Diffeomorphisms between primate skulls

Mathematically, we can describe flexible templates as follows. We must construct four things: (i) a standard image I_T on a domain D_T which can be a set of pixels or can also be reduced to a graph of “parts” of the object, (ii) a space of allowable maps $\psi : D_T \rightarrow D$ or $(D \cup \{\text{missing}\})$, (iii) a measure $E(\psi)$ of the degree of deformation in the map ψ , the “stretching of the springs”, and (iv) a measure of the difference d between the standard image I_T and the deformation $\psi^*(I)$ of the observed image I . Here ψ is typically a diffeomorphism, “missing” is an extra element in the range of ψ to allow certain parts of the standard image to be missing in the observed image, and $\psi^*(I)$ is a “pull-back” of I which may be just the composition of I and ψ if D_T is a set of pixels, or may be some set of local “features” of I when D_T is a graph of parts. The basic algorithm is then to compute

$$\arg \min_{\psi} [d(\psi^*(I), I_T) + E(\psi)], \quad (17)$$

which gives the optimal match of the template with the observed image.

Figures 8b, 8c and 8d show three examples of this algorithm in action. 8b from [Yamamoto-Rosenfeld 82] applies these ideas to the recognition of chinese characters or kanji. In this application D_T is a 1-dimensional polygonal skeleton of the outline of the character, and ψ is a piecewise linear embedding of D_T in the domain D of the character image. 8c from [Y. Amit 91] applies these ideas to tracing a hand in an X-ray by comparing it with a standard hand. Here ψ is a small deformation of the identity defined by a wavelet expansion of its (x, y) -coordinates and the prior $E(\psi)$ is a weighted L^2 -norm of the expansion coefficients. Finally 8d from [Yuille-Hallinan-Cohen 92] applies these ideas to the recognition of eyes. Here D_T has two parts, a pair of parabolas representing the outline of the eye and a black circle on a white ground representing the iris/pupil on the eyeball. ψ is linear on each parabola and on the circle, but the range of the first may *occlude* the range on the second to allow the iris/pupil to be partially hidden. This is incorporated in a careful definition of d .

An interesting mathematical side of this theory is the need for a careful definition and comparative study of various priors on the spaces of diffeomorphisms ψ . One can, for instance, define various measures $E(\psi)$ based (i) on the square norm of the Jacobian, as in harmonic map theory, (ii) on the area of the graph, as in geometric measure theory, (iii) on the stress of the map as in elasticity theory, or (iv) on second derivatives of ψ , which give more control over the minima. [Mumford 91] discusses some of these measures, but the best approach is unclear and restricting maps to be diffeomorphisms is not always natural. An interesting neurophysiological aside is that the anatomy of the cortex of mammals seems well equipped

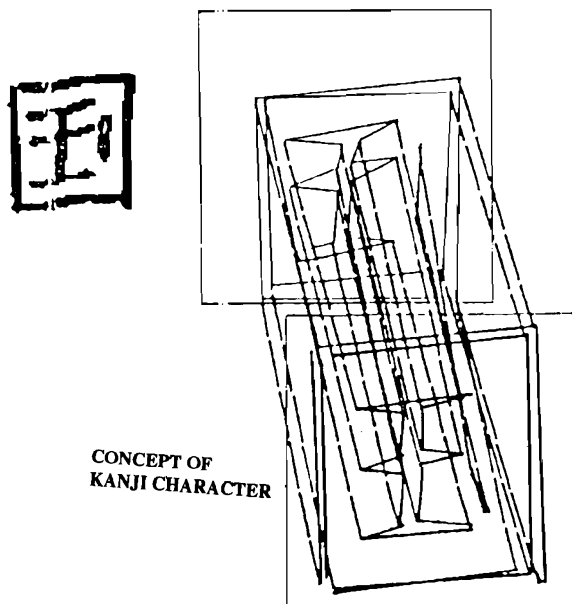


Figure 8b. Diffeomorphism between kanji

to perform domain warping. The circuitry of the cortex is based on two types of connections: local connections within disjoint subsets of the cortex known as *cortical areas*, and global connections, called *pathways*, between the two distinct areas. The pathways occur in pairs, setting up maps which are crudely homeomorphisms between the cortical surfaces of the two areas which are inverse to each other. These pathways are not exactly point to point maps, however, because of the multiple synapses of their axons, hence the pair of inverse pathways can shift a pattern of excitation by small amounts in any direction.

5.4 The stereo correspondence problem via minimum description length

As described in Section 2, there are two approaches to the problems of pattern theory: the first is to use all the geometry, physics, chemistry, biology and sociology that we know about the world and try to define from this high-level knowledge an appropriate probabilistic model ($\Omega_{\text{obs}} \times \Omega_{\text{wld}}, p_{o,w}$) of the world and our observations. The second involves *learning this model* using only the patterns and the internal structure of the signals that are presented to us. Almost all research to date in computer vision falls in the first category, while the standard approach to speech recognition starts with the first but significantly improves on it using the “EM-algorithm”, a learning algorithm in the second category.

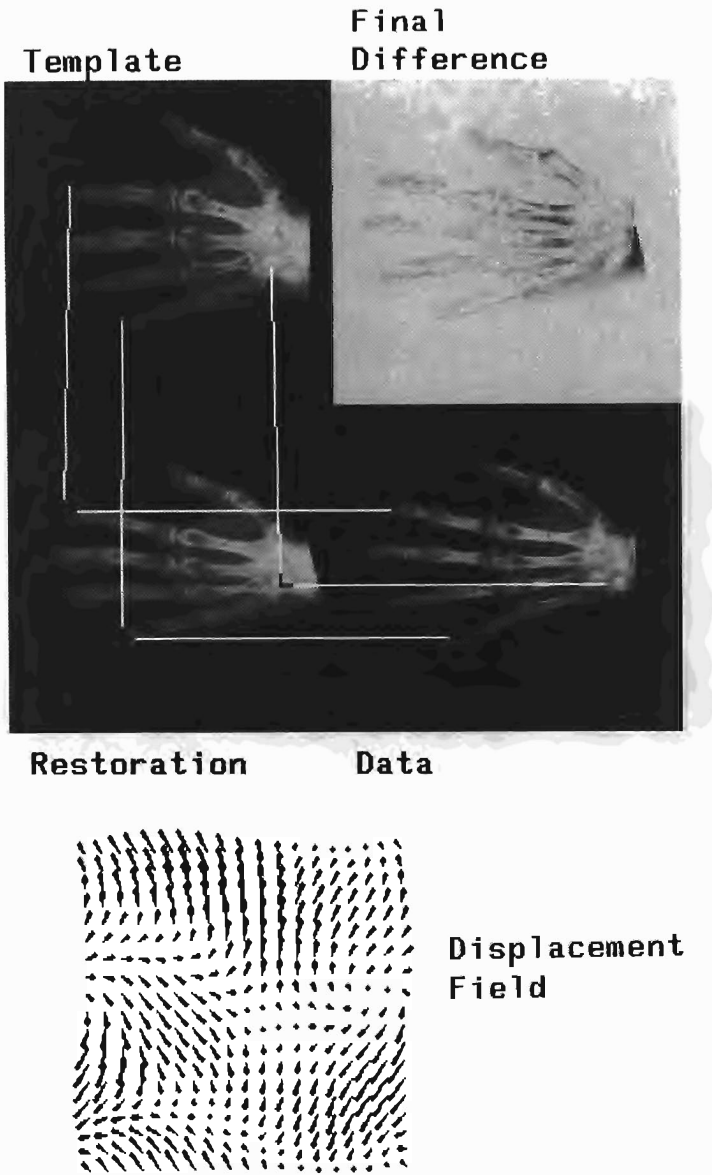


Figure 8c. Diffeomorphism between X-rays of hands

However, newborn animals seem to rely as strongly on learning their environment as on a genetically transmitted knowledge of it. It not hard to imagine that a baby growing up in a virtual reality governed by quite unusual physics would learn these just as rapidly as the physics of its ances-

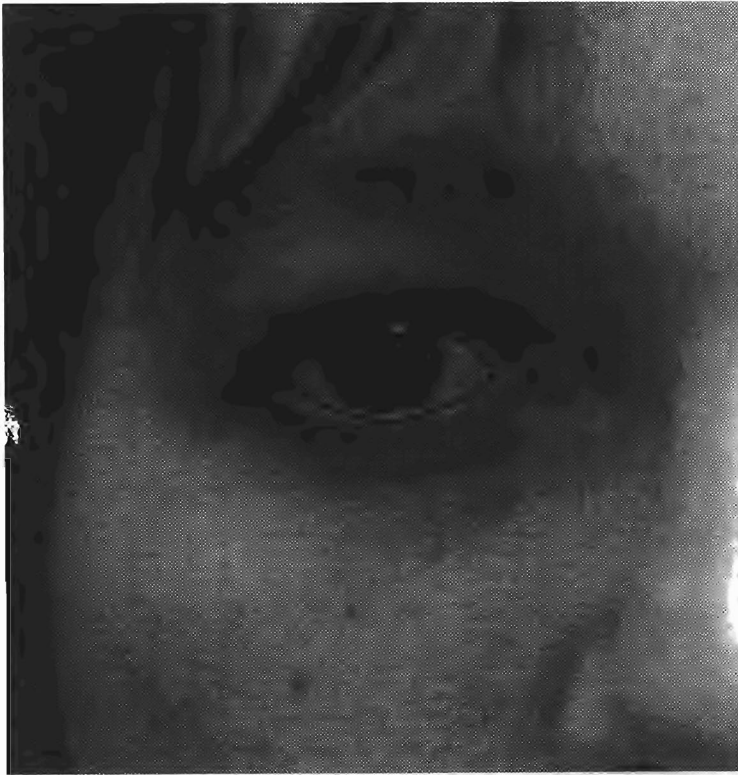


Figure 8d. Diffeomorphism from a cartoon eye to a real eye

tral world. Humans can read scanning electron microscope images, which are produced by totally different reflectance rules from normal images. All of this suggests the possibility of discovering universal pattern analysis algorithms which learn patterns from scratch. One of the great appeals of the idea of pattern theory is the hope that the structure of the world can be discovered in this way. It is in this spirit that we present the final example. It is not an extensive theory like the previous three, but illustrates how the minimum description length principle can lead one to uncover the hidden structure of the world in a remarkably direct way.

We are concerned with the problem of stereo vision. If we view the world with two eyes or with two cameras separated by a known distance, and either identically oriented or with a known difference of orientations, then we can use trigonometry to infer the 3-dimensional structure of the world: see Figure 9. More precisely, the two imaging systems produce two images, I_L and I_R (the left and right images). Suppose a point A in the world visible in both images appears as $A_L \in D_L$ and $A_R \in D_R$ in the domains of the two images. The coordinates of A_L and A_R plus the

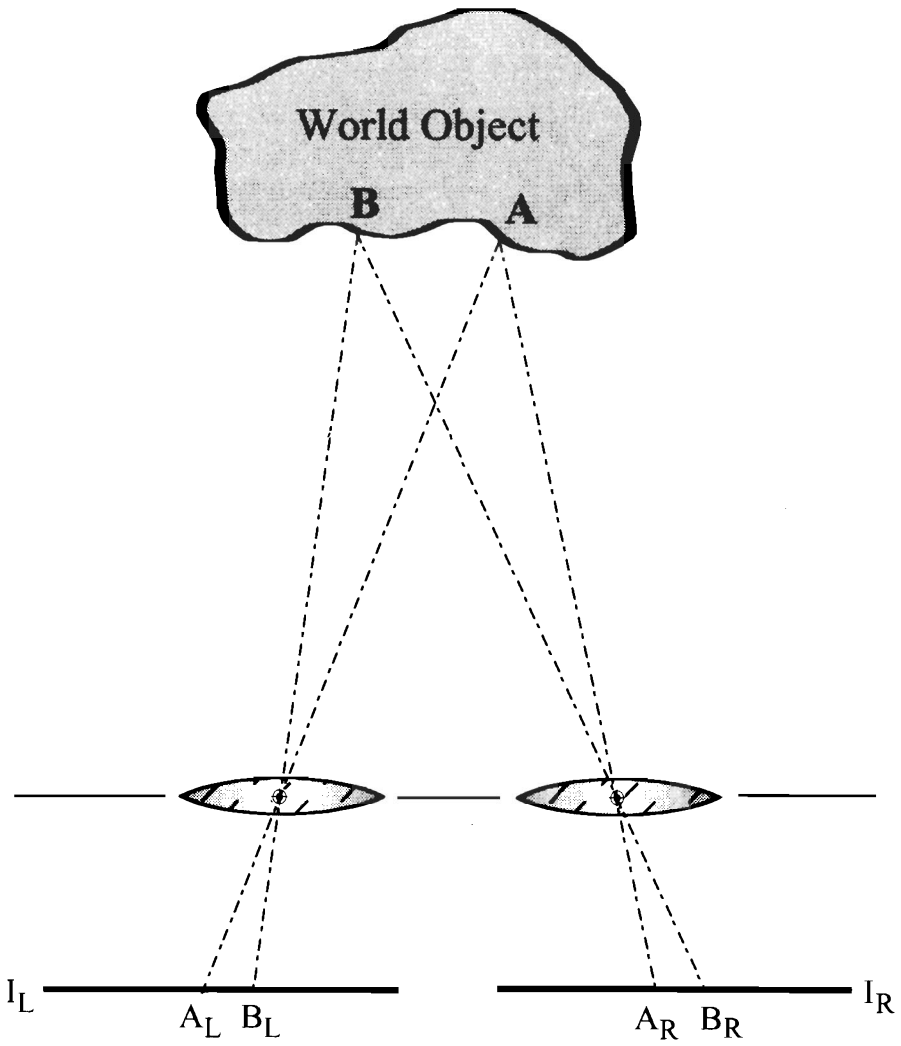


Figure 9. The geometry of stereo vision, in a plane through the centers of the two lenses

known geometry of the imaging system give the 3-dimensional coordinates of A . However, to use this, we need to first find the pair of corresponding points A_L and A_R : finding these is called the *correspondence problem*. Notice from Figure 9 that the geometry of the imaging system gives us one simplification: all points A in a fixed 3-dimensional plane π , through the centers of the two lenses, are seen as points $A_L \in \ell_L$ and $A_R \in \ell_R$, where ℓ_L and ℓ_R are the intersections of π with the two focal planes, and are called *epipolar lines*. Moreover, when we are looking at a single relatively smooth surface S in the 3-dimensional world, S is visible from the left and right eye as subdomains $S_L \subset D_L$ and $S_R \subset D_R$ and the corresponding points on these subdomains define a diffeomorphism $\psi : S_L \rightarrow S_R$ carrying each epipolar line in the left domain to the corresponding epipolar line in the right. This leads us to a problem like that in the last section. But there is a further twist: at the edges of objects, each of the two eyes can typically see a little further around one edge, producing pixels in one domain D_L or D_R with no corresponding pixel in the other domain. In this way, the domain is often segmented into subdomains corresponding to distinct objects.

My claim is that the minimum description length principle alone leads you naturally to discover all this structure, without any prior knowledge of 3-dimensions. The argument is summarized in Figure 10. In this figure, I have represented a series of increasingly complex stereo images in diagrammatic form. Firstly, in order to represent the essentials concisely, I have used only a single pair of epipolar lines ℓ_L and ℓ_R instead of the full domains D_L and D_R . Secondly, instead of graphing the complex intensity function, we have used small symbols (squares, circles, triangles, stars, etc.) to denote local intensity functions with various characteristics. Thus a square on both lines represents local intensities which are similar functions. On the left, at each stage in Figure 10, we see the plane π in the world, with the visible surface points, and the left and right eyes. In the middle, we see the left and right images I_L and I_R which this scene produces, as well as dotted lines connecting corresponding points A_L and A_R . On the right we give a method of encoding the image data.

Stage 0 represents a simple flat object seen from the front: it produces images I_L and I_R , but we assume that our pattern analysis begins with naively encoding the images independently. At stage 1, the same scene is seen, but now the analysis uses the much more concise method of encoding only I_L , the fixed translation d by which corresponding points differ and a possible small residual $\Delta I(x) = I_R(x) - I_L(x + d)$. Clearly this is more concise. At stage 2, the scene is more complex: a surface of varying distance is seen, hence the displacement between corresponding points (called the *disparity*) is not constant. To adapt the previous encoding to this situation,

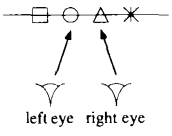
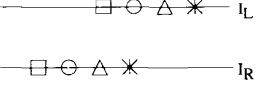
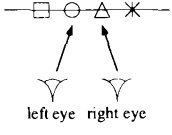
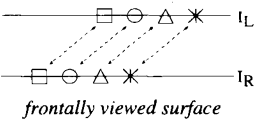
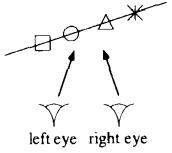
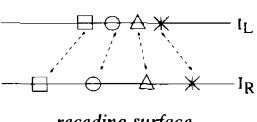
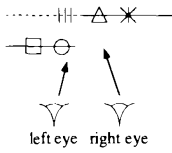
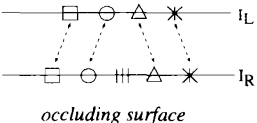
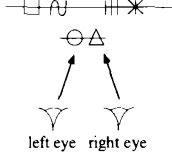
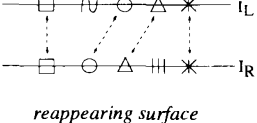
	VIEW OF WORLD (seen from above)	IMAGES I_L and I_R OF WORLD from perspective of left and right eyes	DATA RECORDED
STAGE 0			record raw images I_L and I_R
STAGE 1		 <i>frontally viewed surface</i>	Fit $I_R(x) \sim I_L(x+d)$ - record I_L , d and residual ΔI
STAGE 2		 <i>receding surface</i>	Better fit: $I_R(x) \sim I_L(x+d(x))$ - record I_L , Ave(d), d' , ΔI
STAGE 3		 <i>occluding surface</i>	Still better: $I_C(x) = I_R(x - d(x)/2)$ $\sim I_L(x + d(x)/2)$, where $ d' \leq 1$ - record I_C , Ave(d), d' , ΔI
STAGE 4		 <i>reappearing surface</i>	Best: I_C as above, $d(x)$ from Ave(d_α), d'

Figure 10. Discovering the world via MDL

one could take a mean value of d and have a bigger residual ΔI . But this residual could be quite big and a better scheme is replace the fixed d by a function $d(x)$ and encode I_L , the mean and derivative (\bar{d}, d') of d and the residual ΔI . Now in stage 3, we encounter a new wrinkle: the scene consists in two surfaces, one occluding the other. Notice that a little bit of the back surface is visible to one eye only. To include this complexity, we go over to a more symmetrical treatment of the two eyes and encode a combined *cyclopean* image $I_C(x)$, where

$$I_C(x) = I_R(x - \frac{d(x)}{2}), I_L(x + \frac{d(x)}{2}) \text{ or their average} \quad (18)$$

depending on whether the point is visible only to the right eye, only to the left eye or to both eyes. To make this representation unique, it is easy to see that we must require that $|d'(x)| \leq 1$. Then we encode the scene via $(I_C, \bar{d}, d', \Delta I)$. In the final stage 4, we introduce the possibility of a surface disappearing behind another *and then reappearing*. The point is that each surface has its own average disparity, and it now becomes more efficient to record d by several means \bar{d}_α , one for each surface, and the derivative d' . Thus we see how the search for minimum length encoding leads us naturally, first to the third coordinate of world points, then to smooth descriptions of surfaces in terms of their tangent planes and finally to explicit labelling of distinct surfaces in the visible field.

Although this approach might seem very abstract and impossible to implement biologically, G. Hinton (unpublished) has developed neural net theories incorporating both MDL and feed-back. These might be able to learn stereo exactly as outlined in this section.

6. Pattern theory and cognitive information processing

The examples of the last section all concern pattern theory as a theory for analyzing sensory input. The examples come from vision, but most of the ideas could apply to hearing or touch too. The purpose of this section is to ask the question: to what extent is pattern theory relevant to all cognitive information processing, both "higher level" thinking as studied in cognitive psychology and AI, and the output stages of an intelligent agent, motor control and action planning. I believe that in many ways the same ideas are applicable on a theoretical level and that there is physiological evidence that the same algorithms are applied throughout the cortex.

In the introduction, we gave medical expert systems as another example of pattern theory. In this extension, we considered the data available to a physician — symptoms, test results and the patient's history — as

an encoded version of the full state of the world, a “deformed image” in Grenander’s terminology. The full state of the world, the “pure image”, in this case means the diseases and processes present in the patient. Inferring these hidden random variables can and has been studied as a problem in Bayesian statistics, exactly as in Section 2: see, for instance, [Pearl 88], [Lauritzen-Spiegelhalter 88]. In particular, describing the probability distribution on all the random variables as a Gibbs field, as in Section 5b, has been shown to be a powerful technique for introducing realistic models of the probability distribution in the real world. Figure 11, from the article [Lauritzen-Spiegelhalter 88], shows a simplified set of such random variables and the graph on which a Gibbs distribution can be based.

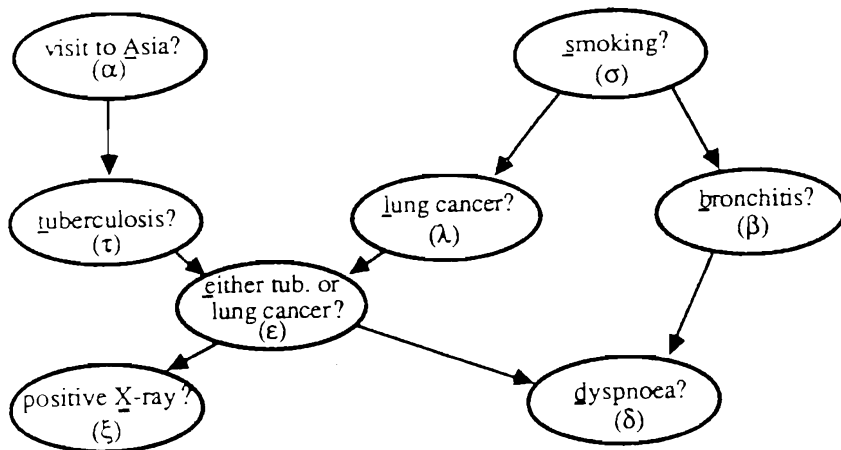


Figure 11. Causal graph in a toy medical expert system

Whether or not pattern theory extends in an essential way to these types of problems hinges on whether the transformations described in Section 3 generate the kind of probability distributions encountered with higher level variables. To answer this, it is essential to look at test cases which are not too artificially simplified (as is done all too often in AI), but which incorporate the typical sorts of complexities and complications of the real world. While I do not think this question can be definitively answered at present, I want to make a case that the four types of transformations of Section 3 are indeed encoding mechanisms encountered at all levels of cognitive information processing.

The first class of transformations, noise and blur, certainly occur at all levels of thought. In the medical example, errors in tests, the inadequacies

of language in conveying the nature of a pain or symptom, etc. all belong to this category. Uncertainty over facts, misinterpretations and confusing factors are within this class. The simplest model leads to multi-dimensional normal distributions on a vector P of “features” being analyzed.

The fourth category of transformation, “interruptions”, also are obviously universal. In any cognitive sphere, the problem of separating the relevant factors for a specific event or situation being analyzed from the extraneous factors involved with everything else in the world, is clearly central. The world is a complex place with many, many things happening simultaneously, and highlighting the “figure” against the “ground” is not just a sensory problem, but one encountered at every level. Another way this problem crops up is that a complex of symptoms may result from one underlying cause or from several, and, if several causes are present, their effects have to be teased apart in the process of pattern analysis. As proposed in Section 4, pattern synthesis — actively comparing the results of one cause with the presenting symptoms P followed by analysis of the residual, the unexplained symptoms, is a universal algorithmic approach to these problems.

The second of the transformations, “multi-scale superposition”, can be applied to higher level variables as follows: philosophers, psychologists and AI researchers have all proposed systematizing the study of concepts and categories by organizing them in hierarchies. Thus psychologists (see [Rosch 78]) propose distinguishing *superordinate categories*, *basic level categories* and *subordinate categories*: for instance, a particular pet might belong to the superordinate category “animal”, the basic-level category “dog” and the subordinate category “terrier”. In AI, this leads to graphical structures called *semantic nets* for codifying the relationships between categories (see [Findler 79]). These nets always include ordered links between categories, called *isa* links, meaning that one category is a special case of another. I want to propose that cognitive multi-scale superposition is precisely the fact that to analyze a specific situation or thing, some aspects result from the situation belonging to very general categories, others from very specific facts about the situation that put it in very precise categories. Thus sensory thinking requires we deal with large shapes with various overall properties, supplemented with details about their various parts, precise data on location, proportions, etc.; cognitive thinking requires we deal with large ideas with various general properties, supplemented with details about specific aspects, precise facts about occurrence, relationships, etc.

Finally, how about “domain warping”? Consider a specific example first. Associated to a cold is a variety of several dozen related symptoms.

A person may, however, be described as having a sore throat, a chest cold, flu, etc.: in each case the profile of their symptoms shifts. This may be modelled by a map from symptom to symptom, carrying for instance the modal symptom of soreness of throat to that of coughing. The more general cognitive process captured by domain warping is that of making an *analogy*. In an analogy, one situation with a set of participants in a specific relationship to each other is mapped to another situation with new participants in the same relationship. This map is the ψ in Section 5c, and the constraints on ψ , such as being a diffeomorphism, are now that it preserve the appropriate relationships. The idea of domain warping applying to cognitive concepts seems to suggest that higher level concepts should form some kind of geometric space. At first this sounds crazy, but it should be remembered that the entire cortex, high and low level areas alike, has the structure of a 2-dimensional sheet. This 2-dimensional structure is used in a multitude of ways to organize sensory and motor processes efficiently: in some cases, sensory maps (like the retinal response and patterns of tactile responses) are laid out geometrically. In other cases, interleaved stripes carry intra-hemispheric and inter-hemispheric connections. In still other cases, there are “blobs” in which related responses cluster. But, in all cases, adjacency in this 2D sheet allows a larger degree of cross-talk and interaction than with non-adjacent areas and this seems to be used to develop responses to related patterns. My suggestion is: is this spatial adjacency used to structure abstract thought too*?

To conclude, we want to discuss briefly how pattern theory helps the analysis of motor control and action planning, the output stage of a robot. Control theory has long been recognized as the major mathematical tool for analyzing these problems but it is not, in fact, all that different from pattern theory. In Figure 12a, we give the customary diagram of what control theory does. The controller is a black box which compares the

* I have argued elsewhere that the remarkable anatomical uniformity of the neo-cortex suggests that common mechanisms, such as the 4 universal transformations of pattern theory, are used throughout the cortex [Mumford 91-92, 93]. The referee has pointed out that “the uniformity of structure may reflect common machinery at a lower level. For example, different computers may have similar basic mechanisms at the level of registers, buses, etc., which is a low level of data handling. Similarly in the brain, the apparent uniformity of structure may be at the level of common lower-level mechanisms rather than the level of dealing with universal transformations”. This is a certainly an alternative possibility, quite opposite to my conjectural link between the high-level analysis of pattern theory and the circuitry of the neo-cortex.

current observation of the state of the world with the desired state and issues an updated motor command, which in turn affects the black box called the world.

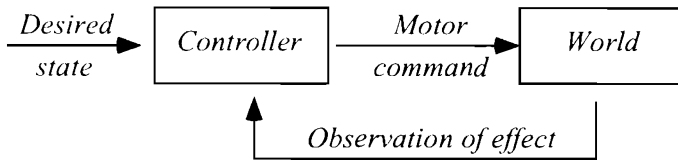


Figure 12a. The flow chart of control theory

This diagram is very similar to Figure 4, which described how pattern analysis and pattern synthesis formed a loop used in the algorithm for reconstructing the hidden world variables from the observed sensory ones.

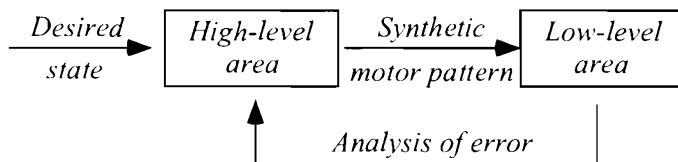


Figure 12b. A motor task via pattern theory

Figure 12b presents the modification of Figure 4 to a motor task. Here a high-level area or “black box” is in a loop with a low-level area: the high-level area compares the desired state with an analysis of the error and generates an updated motor command sequence by pattern synthesis. The low-level area, either by actually carrying out an action and observing its consequences, or by internal simulation, finds that it falls short in various ways, and send its pattern analysis of this error back up. Notice that the four transformations of Section 3 will occur or should be used in the top-down pattern synthesis step. Noise and blur are the inevitable consequences of the inability to control muscles perfectly, or eliminate external uncontrollable interference. Domain warping is the bread-and-butter of control theory — speeding up or slowing down an action by modifying the forces in order that it optimizes performance. Multi-scale superposition is what hierarchical control is all about: building up an action first in large steps, then refining these steps in their parts, eventually leading to

detailed motor commands. Finally, interruptions are the terminations of specific control programs, either by success or by unexpected events, where quite new programs take over. In general, we seek to anticipate these and set up successor programs beforehand, hence we need to actively synthesize these as much as possible.

In summary, my belief is that pattern theory contains the germs of a universal theory of thought itself, one which stands in opposition to the accepted analysis of thought in terms of logic. The successes to date of the theory are certainly insufficient to justify such a grandiose dream, but no other theory has been more successful. The extraordinary similarity of the structure of all parts of the human cortex to each other and of human cortex with the cortex of the most primitive mammals suggests that a relatively simple universal principle governs its operation, even in complex and deep thinking (see [Mumford 91-92, 93] where these physiological links are developed).

Bibliography

- Ambrosio, L. and Tortorelli, V., Approximations of functionals depending on jumps by elliptic functionals via gamma-convergence, *Comm. Pure & Applied Math.* **43** (1991), 999-1036.
- Amit, D., *Modelling Brain Function*, Cambridge University Press, Cambridge, 1989.
- Amit, Y., A non-linear variational problem for image matching, *SIAM Journal on Scientific Computing*, to appear.
- Belhumeur, P. and Mumford, D., A Bayesian Treatment of the Stereo Correspondence Problem Using Half-Occluded Regions, in *Proc. IEEE Conf. Comp. Vision and Pattern Rec.*, 1992, 506-512.
- Blake, A. and Zisserman, A., *Visual Reconstruction*, MIT Press, Cambridge, 1987.
- Burt, P. and Adelson, E., The Laplacian pyramid as a compact image code, *IEEE Trans. on Comm.* **31** (1983), 532-540.
- Carpenter, G. and Grossberg, S., A massively parallel architecture for a self-organizing neural pattern recognition machine, *Comp. Vision, Graphics and Image Proc.* **37** (1987), 54-115.
- Cavanagh, P., What's up in top-down processing, *Proc. 13th ECVP*, 1991.
- Coifman, R., Meyer, Y. and Wickerhouser, V., *Wavelet analysis and signal processing*, preprint, Yale Univ. Math. Dept., New Haven, 1990.

- Dal Maso, G., Morel, J-M. and Solimini, S., A Variational Method in Image Segmentation, *Acta Math.* **168** (1992), 89–151.
- Daubechies, I., Orthonormal bases of compactly supported wavelets, *Comm. Pure & Applied Math.* **49** (1988), 909-996.
- Daubechies, I., The wavelet transform, time-frequency localization and signal analysis, *IEEE Trans. Inf. Theory*, 1990, 961-1005.
- DeGiorgi, E., Carriero, M. and Leaci, A., Existence theorem for a minimum problem with free discontinuity set, *Arch. Rat. Mech. Anal.* **108** (1989), 195-218.
- Dowling, J., *The Retina*, Harvard University Press, Cambridge, 1987.
- Findler, N. ed., *Associative Networks*, Academic Press, New York, 1979.
- Fischler, M. and Elschlager, R., The Representation and Matching of Pictorial Structures, in *IEEE Trans. on Computers* **22** (1973), 67-92.
- Gage, M. and Hamilton, R., The heat equation shrinking convex plane curves, *J. Diff. Geom.* **23** (1986), 69-96.
- Geiger, D. and Yuille, A., A Common Framework for Image Segmentation, *Int. J. Comp. Vision*, 1990.
- Geiger, D., Ladendorf, B. and Yuille, A., Occlusions and Binocular Stereo, in *Proc. European Conf. Comp. Vision*, **588** (1992), Springer Lecture Notes in Computer Sciences, Berlin-Heidelberg.
- Geman, D., *Random Fields and Inverse Problems in Imaging*, Springer Lecture Notes in Math. **1427** (1991), Berlin-Heidelberg-New York.
- Geman, S. and Geman, D., Stochastic relaxation, Gibbs distribution and Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **6** (1984), 721-741.
- Geman, S., Geman, D., Graffigne, C. and Dong, P., Boundary detection by Constrained Optimization, *IEEE Trans. Pattern Anal. and Mach. Int.* **12** (1990).
- Grayson, M., The heat equation shrinks embedded plane curves to round points, *J. Diff. Geom.* **26** (1987), 285-314.
- Grenander, U., *Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern Synthesis and Regular Structures*, Springer-Verlag, Heidelberg-New York, 1976-1981.
- Grossberg, S. and Mingolla, E., Neural dynamics of form perception: boundary completion, illusory figures and neon color spreading, *Psych. Rev.* **92** (1985), 173-211.

Grossman, A. and Morlet, J., Decomposition of Hardy functions into square integrable wavelets of constant shape, *SIAM J. Math. Anal.* **15** (1984), 723-736.

Hertz, J., Krogh, A. and Palmer, R., *Introduction to the Theory of Neural Computation*, Addison-Wesley, 1991.

Hildreth, E., *The Measurement of Visual Motion*, MIT Press, Cambridge, 1984.

Hopfield, J., Neural Networks and Physical Systems with Emergent Collective Computational Abilities, in *Proc. Nat. Acad. Sci.* **79**(1982), 2554-2558.

Kass, M., Witkin, A. and Terzopoulos, D., Snakes: Active Contour Models, *IEEE Proc. 1st Int. Conf. Computer Vision*, 1987, 259-268.

Kirkpatrick, S., Geloti, C. and Vecchi, M., Optimization by Simulated Annealing *Science* **220** (1983), 671-680.

Lauritzen, S. and Spiegelhalter, D., Local Computations with Probabilities on Graphical Structures and their Applications to Expert Systems, *J. Royal Stat. Soc. B* **50** (1988), 157-224.

Leclerc, Y., Constructing simple stable descriptions for image partitioning, *Int. J. Comp. Vision* **3** (1989), 73-102.

Lee, T.S., Mumford, D. and Yuille, A., Texture Segmentation by Minimizing Vector-Valued Energy Functionals, in *Proc. Eur. Conf. Comp. Vision*, Springer Lecture Notes in Computer Science **1427** (1992).

Mallat, S., A theory of multi-resolution signal decomposition: the wavelet representation, *IEEE Trans. PAMI* **11** (1989), 674-693.

Meyer, Y., Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs, *Séminaire Bourbaki*, Springer Lecture Notes in Math, Berlin-Heidelberg-New York, (1986).

Mumford, D., Mathematical theories of shape: Do they model perception?, *Proc. SPIE* **1570** (1991), 2-10.

Mumford, D., On the computational architecture of the neocortex I and II, *Biol. Cybernetics* **65** (1991-92), 135-145 & 66, 241-251.

Mumford, D., Neuronal architectures for pattern-theoretic problems, in *Proc. Idyllwild conference on large scale neuronal theories of the brain*, C. Koche (ed.), MIT Press, Cambridge, 1994.

Mumford, D. and Shah, J., Optional approximation by piecewise smooth functions and associated variational problems, *Comm. Pure & Applied Math.* **42** (1989), 577-685.

- Nitzberg, M. and Mumford, D., The 2.1D sketch, in *Proc. 3rd IEEE Int. Conf. Comp. Vision*, 1990, 138-144.
- Osherson, D. and Weinstein, S., Formal Learning Theory, in *Handbook of Cognitive Neuroscience* M. Gazzaniga (ed.), Plenum Press, New York, 1984, 275-292.
- Pearl, J., *Probabilistic Reasoning in Intelligent Systems*, Morgan-Kaufman, 1988.
- Perona, P. and Malik, J., Scale-space and edge detection using anisotropic diffusion, *IEEE Workshop on Computer Vision*, Miami, 1987.
- Rissanen, J., *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.
- Rosch, E., Principles of Categorization, in *Cognition and Categorization*, E. Rosch and B. Lloyd (eds.), L. Erlbaum, 1978.
- Rosenfeld, A. and Thurston, M., Edge and curve detection for visual scene analysis, *IEEE Trans. on Computers* **C-20** (1971), 562-569.
- Thompson, D'Arcy, *On Growth and Form*, Cambridge University Press, Cambridge, 1917.
- Uhr, L., Layered "recognition cone" networks that preprocess, classify and describe, *IEEE Trans. on Computers* **C-21** (1972), 758-768.
- Yamamoto, K. and Rosenfeld, A., Recognition of Handprinted Kanji Characters by a Relaxation Method, in *Proc. 6th Int. Conf. Pattern Recognition*, 1982, 395-398.
- Yuille, A. and Grzywacz, N., A Mathematical Analysis of the Motion Coherence Theory, *Int. J. Comp. Vision* **3**(1989), 155-175.
- Yuille, A., Hallinan, P., and Cohen, D. Feature Extraction from Faces using Deformable Templates, *Int. J. Comp. Vision* **6** (1992).

Department of Mathematics
Harvard University
Cambridge, MA 02138, USA

Received September 28, 1992