Pattern Theory through Examples

David Mumford and Agnes Desolneux

December 12, 2005

Preface: the nature of this book

The title of the book *Pattern Theory through Examples* was chosen to put up front the fact that we are trying to write a basic introductory text from a non-standard point of view. Our goal is not to systematically expound some area of mathematics as one would do in a graduate course in mathematics. Nor is it to explain how to develop all the basic tools for analyzing speech signals or images as one would do in a graduate course in engineering. In our view, what makes applied mathematics distinctive is that one starts with a collection of problems from some area of science and one seeks the appropriate mathematics for clarifying the experimental data and the underlying processes producing this data. One needs mathematical tools and, almost always, computational tools as well. The practitioner finds his or herself engaged in a lifelong dialog: seeking models, testing them against data, identifying what is missing in the model and refining the model. In some situations, the challenge is to find the right mathematical tools; in others, it is to find ways of efficiently calculating the consequences of the model, as well as finding the right parameters in the model. This book, then, seeks to actively bring the reader into such a dialog.

To do this, each Chapter begins with some example, a class of signals whose variability and whose patterns we seek to model. Typically, some mathematical tools are called for. We put these in sections labelled **Basics I,II, ...** and in these sections we put on our mathematical hats and simply develop the math. In most cases, we give the basic definitions, state some results and prove a few of them. But we refer the reader elsewhere for a full exposition of the topic. Then we compare the model with the data. Usually, new twists emerge. At some point, questions of computability emerge. Here we also have special sections labelled **Algorithm I,II,...**. Each Chapter ends with a set of exercises. Both of the authors have taught some or part of this material multiple times and have drawn on these classes. In fact, the book is based on lectures delivered in 1998 at the Institut Henri Poincaré by the senior author and the notes made at the time by the junior author.

The Chapters are ordered by the complexity of the signal being studied. In Chapter 1, we will look at text strings: discrete valued functions in discrete time. In Chapter 2, we will look at music, a real valued function of continuous time. In Chapter 3, we will look at character recognition, this being mostly about curvilinear structures in the domain of a function of 2 variables (viz. an image $I(x, y)$). In Chapter 4, we deal with the separation of textures in an image, a fully 2-dimensional aspect of images. In Chapter 5, we deal with faces, where diffeomorphisms of planar domains are the key player. In Chapter 5, we deal with scaling effects present in natural images, their statistical self-similarity.

There is a common thread in all these topics. The Introduction which follows explains this thread, the approach to the analysis of signals which we call Pattern Theory. In the choice of specific algorithms and ideas, the authors have, of course, used the research they know best. Hence the work of their own groups, their students and collaborators, is heavily represented. Nonetheless, they have tried to present a fair cross-section, especially, of computer vision research.

Contents (not yet finalized)

Notations used throughout the book

$P, Q, ...$ will be probability distributions on a countable set, always given by upper case letters. Their values are $P(x), Q(y), ...$

$P, Q, ...$ as well as $\mu, \nu, ...$ will be probability measures on continuous spaces. Their values on measurable subsets will be $P(A), \mu(B), ...$ *If* there is a reference measure dx, their probability densities will be denoted by lower case letters: $P(dx) = p(x)dx$ or $dP/dx = p(x)$. Thus we write:

$$P(A) = \int_A dP(x) = \int_A p(x)dx.$$

$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{a}$, ... will be random variables from the set for the given probability distribution. Note that random variables are always indicated by this distinctive font.

$\mathbb{P}$(predicate with random variables) will be the probability of the predicate being true. If needed, we write $\mathbb{P}_P$ to indicate the probability w.r.t. P. Thus

$$P(x) = \mathbb{P}(\mathcal{X} = x) = \mathbb{P}_P(\mathcal{X} = x)$$
$$P(A) = \mathbb{P}(\mathcal{X} \in A) = \mathbb{P}_P(\mathcal{X} \in A).$$

$\mathbb{E}(f)$ will be the expectation of the random variable or function of random variables f. If needed, we write $\mathbb{E}_P$ to indicate the expectation w.r.t. P. f, for instance, can be $\log(P(\mathcal{X}))$ which we abreviate to $\log(P)$.

Introduction: What is Pattern Theory?

The term ‘Pattern Theory’ was coined by Ulf Grenander to distinguish his approach to the analysis of patterned structures in the world from ‘pattern recognition’. In this book, we use it in a rather broad sense to include the statistical methods used in analyzing all ‘signals’ generated by the world, whether they be images, sounds, written text, DNA or protein strings, spike trains in neurons, time series of prices or weather; examples from all of these appear either in Grenander’s book ‘Elements of Pattern Theory’ or in the work of our colleagues, collaborators and students on Pattern Theory. We believe the work in all these areas has a natural unity: common techniques and motivations. In particular, Pattern theory proposes that the types of patterns (and the hidden variables needed to describe these patterns) which are found in one class of signals will often be found in the others and that their characteristic variability will be similar. Hence the stochastic models used to describe signals in one field will crop up in all the others. The underlying idea is to find classes of stochastic models which can capture all the patterns which we see in nature, so that random samples from these models have the same ‘look and feel’ as the samples from the world itself. Then the detection of patterns in noisy and ambiguous samples can be achieved by the use of Bayes’s rule, a method which can be described as ‘analysis by synthesis’.

1. The manifesto of pattern theory

We want to express this approach to signals and their patterns in a set of five basic principles. The first three are:

1 A wide variety of signals result from observing the world, all of which show patterns of many kinds, which are caused by objects, processes and laws present in the world but at least partially hidden from direct observation. These patterns can be used to infer information about these unobserved factors.

2 Observations are affected by many variables which are not conveniently modeled deterministically because they are too complex or too hard to observe and often belong to other categories of events which are irrelevant to the observations of interest. To make inferences in real time or with a model of reasonable size, we must model our observations partly stochastically and partly deterministically.

3 Accurate stochastic models which capture the patterns present in the signal, while respecting their natural structures, i.e. symmetries, independences of parts, marginals on key statistics, are needed. These models should be learnt from the data and validated by sampling: inferences from them can be made using Bayes's rule provided that samples from them resemble real signals.

To fix ideas, a microphone or an ear observes the pressure wave $p(t)$ transmitted by the air from a speaker's mouth. In the figure, this function $p(t)$ has many obvious patterns, e.g. it divides into four distinct segments in which it has very different character. These are four 'phones' caused by changes in the configuration of the mouth and vocal cords of the speaker during the passage of air, which in turn are caused by the intention in the speaker's brain to utter a certain word. These patterns in p encode in a noisy, highly variable way the sequence of phones being pronounced and the word which these phones make up. We cannot *observe* the phones or the word directly, hence they are called hidden variables, but must infer them. Early work on speech recognition attempted to make this inference deterministically using logical rules based on binary features extracted from $p(t)$. For instance, the table below shows some of the features used to distinguish english consonants (NEED INT. PHONETIC SYMBOLS):

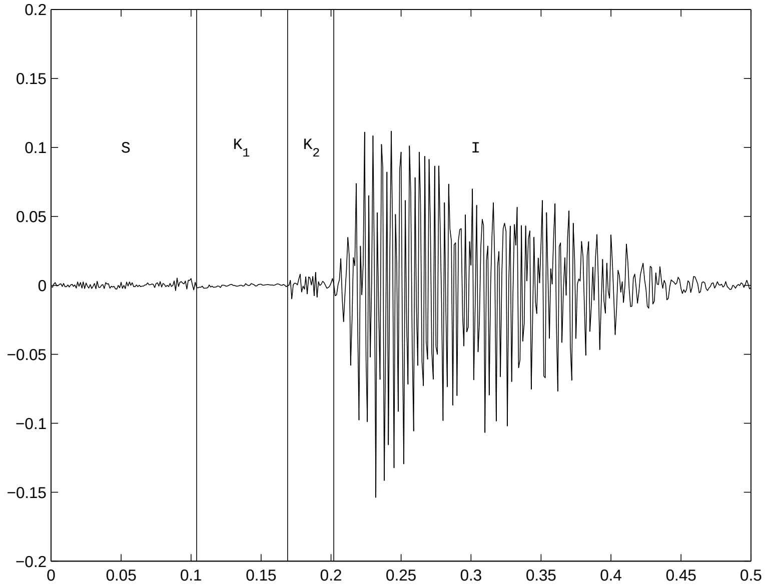

Figure 1: Half a second of the raw acoustic signal during the pronounciation of the word 'ski'. Note the 4 phones: (i) low intensity white noise during the letter 's', (ii) silence while the mouth is closed during the beginning of 'k', (iii) a burst when the mouth opens and the rest of 'k' is pronounced and (iv) a prolonged harmonic sound for the vowel 'i'.

	p	t	k	b	d	g	m	n	ng	f	th	s	sh	v	dh	z	zh	ch	jh
Continuant	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	−	−
Voiced	−	−	−	+	+	+	+	+	+	−	−	−	−	+	+	+	+	−	+
Nasal	−	−	−	−	−	−	+	+	+	−	−	−	−	−	−	−	−	−	−
Labial	+	−	−	+	−	−	+	−	−	+	−	−	−	+	−	−	−	−	−
Coronal	−	+	−	−	+	−	−	+	−	−	+	+	+	−	+	+	+	+	+
Anterior	+	+	−	+	+	−	+	+	−	+	+	+	−	+	+	+	−	−	−
Strident	−	−	−	−	−	−	−	−	−	+	−	+	+	+	−	+	+	+	+

This deterministic rule-based approach failed and the state-of-the-art is to use a family of precisely tuned stochastic models (Hidden Markov Models) and a Bayesian MAP estimator to do better.

The identical story played out in vision, in parsing sentences from text, in expert systems, etc. In all cases, the initial hope was that the deterministic laws of physics plus logical syllogisms for combining facts would give reliable methods for decoding the signals of the world in all modalities. These naive approaches seem to always fail because the signals are too variable and the hidden variables too subtly encoded. One reason is that there are always so many extraneous factors effecting the signal: noise in the microphone and the ear, other sounds in the room besides the speaker, variations in the geometry of the speaker's mouth, the speaker's mood, etc. Although, in principle, one might hope to know these as well as the word being pronounced, inferring each extra factor makes the task harder. (In some cases, like radioactive decay in PET scans, it is even impossible by quantum theory.) Thus stochastic models are *required*.

Now the world is very complex and the signals it generates are likewise complex both in their patterns and in their variability. If pattern theory makes any sense, there must be some rules whereby the patterns caused by the objects, processes and laws in the world are not like some arbitrary recursively enumerable set. For instance, in the movie 'Contact' Jodie Foster discovers life in outer space because it broadcasts a signal encoding the primes from 2 to 101. If an infant had to recognize this sort of pattern in order to learn to speak, they would never succeed. By and large, the patterns in the signals received by our senses are correctly learned by infants, at least to the level required to reconstruct the various types of objects in our world, their properties and to communicate with adults. This is a marvelous fact and pattern theory attempts to understand why this is so. Two further key ideas can be added to the manifesto to explain this:

4 The various objects, processes and rules of the world produce patterns which can be described as precise *pure patterns* distorted and transformed by a limited family of *deformations*, similar across all modalities.

5 When all the stochastic factors affecting any given observation are suitably identified, they show a large amount of conditional independence.

To do pattern theory properly, it is essential to identify the patterns present in the signal correctly. We often have an intuitive idea of what are the important patterns, but the human brain does many things unconsciously and also takes lots of short-cuts to get things done fast. Thus a careful analysis of the actual data to see what the data is telling us is preferable to slapping together an 'off the shelf' Gaussian or log-linear model based on our guesses. Here is a very stringent test of whether a stochastic model is a good desciption of the world: *sample from it.* This is so obvious that one would assume everyone does this. But actually, this is not so. The samples from many models that are used in practice are absurd oversimplifications of real signals and, even worse, some theories do not include the signal itself as one of its random variables (using only some derived variables), so it is not even possible to sample signals from them[1].

In what ways is Pattern Theory different from the better known field of Statistical Pattern Recognition? Traditionally, the focus of statistical pattern recognition was the study of one or more data sets $\{\vec{x}_\alpha \in \mathbb{R}\}_{\alpha \in I}$ with the goal of (a) fitting (parametric and non-parametric) probability distributions to each data set, (b) finding optimal decision rules for classifying new data into the correct data set and (c) separating a single data set into clusters when it appears to be a mixture. The essential issue is the 'bias-variance' trade off: to model fully the complexities of the data source but not the accidental variations of the specific data set. When Grenander first proposed Pattern Theory as a distinct enterprise, there were several very novel aspects of his approach:

- Firstly, he proposed that to describe the patterns in typical datasets, one should always look for appropriate hidden variables, in terms of which the patterns were more clearly described.
- Secondly, he proposed that the set of variables, observed and hidden, typically formed the vertices of a graph, as in Gibbs models, and that one must formulate prior probability distributions for the hidden variables as well as models for the observed variables.
- Thirdly, he proposed that this graph itself might be random and its variability must then be modeled.
- Fourthly, he proposed that one could list the different types of 'deformations' which patterns were subject to, thus creating the basic classes of stochastic models that can be applied.
- Fifthly, he proposed that these models should be used for pattern synthesis as well as analysis.

As the subject evolved, statistical pattern recognition merged with the area of neural nets and the first two ideas were absorbed into statistical pattern recognition. So called 'graphical models' are now seen as the bread and butter of the field and discovering these hidden variables is a challenging new problem. And the use of prior models has become the mainstream approach in vision and expert systems as it has been in speech since the 60's. However, the other aspects of Pattern Theory are still quite distinctive.

2. The basic types of patterns

I want to be more precise about the kinds of patterns and deformations referred to point 4 above. Real world signals show two very distinct types of patterns. I want to call these: i) *value patterns*

[1] This was, for instnce, the way most traditional speech recognition systems worked: their approach was to throw away the raw speech signal in the pre-processing stage and replace it with codes designed ignore speaker variation. In contrast, in our own speech recognition, we are clearly aware of the individual speaker's type of voice and of any departures from normal.

and ii)*geometrical patterns*. Signals, in general, are some sort of functions $f : X \to V$. The domain may be continuous (e.g. a part of space like the retina or an interval of time) or discrete (e.g. the nodes of a graph or a discrete sample in space or time) and the range may be a vector space or binary $\{0, 1\}$ or something in the middle. In the case of value patterns, we mean that the features of this pattern are computed from linear combinations of the values of f (for example, power in some frequency band). In the case of geometric patterns, the function f can be thought of as producing geometrical patterns in its domain (for example, the set of its points of discontinuity). The distinction affects which extra random variables you need to describe the pattern. For value patterns, we typically add coefficients in some expansion in order to describe the particular signal. For geometric patterns, we add certain points or subsets of the domain or features of such subsets. Traditional statistical pattern recognition and the traditional theory of stationary processes deals only with the values of f, not the geometry of its domain. Let me be more specific and describe these patterns more explicitly. And I want to go further and distinguish two sorts of geometric patterns: those involving one geometry which can be deformed and those involving a hierarchical geometric pattern:

1. *Value patterns and linear superposition.* The most standard value model creates the observed signal from the linear superposition of fixed or learned basis functions. We assume the observed signal has values in a real vector space: $s : X \longrightarrow V$ and that it is expanded in terms of auxiliary functions $s_\alpha : X \longrightarrow V$ as:

   $$s = \Sigma_\alpha c_\alpha s_\alpha.$$

   In vision, X is the set of pixels and, for a greylevel image, V is the set of real numbers. Here the coefficients $\{c_\alpha\}$ are random variables while the functions $\{s_\alpha\}$ may be either a universal basis like sines and cosines or wavelets or some learned templates as in Karhunen-Loeve expansions. Or the $\{s_\alpha\}$ may be random: the simplest case is to allow one of them to be random and think of it as the residual, e.g. an additive noise term. An important case is that of expanding a function into its components on various scales (as in wavelet expansions), so the terms s_α are simply the components of s on scale α. See figure 2 for an example where a face is expanded into three images representing its structure on fine, medium and coarse scales. Other variants are i) amplitude modulation in which the low and high frequencies are combined by multiplication instead of addition and ii) the case where s is just a discrete sample from the full function $\Sigma c_\alpha s_\alpha$. Quite often, the goal of such an expansion from a statistical point of view is to make the coefficients $\{c_\alpha\}$ as independent as possible. In case the coefficients are Gaussian and independent, this is called PCA or a Karhunen-Loeve expansion; in the non-Gaussian but independent case, it is called ICA. We shall see how such expansions work very well to express the effects of lighting variation on face signals.

2. *Simple geometric patterns and 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’. Assume that the observed signal is a random map $s : X \longrightarrow V$ as above. Then our model includes the warping which

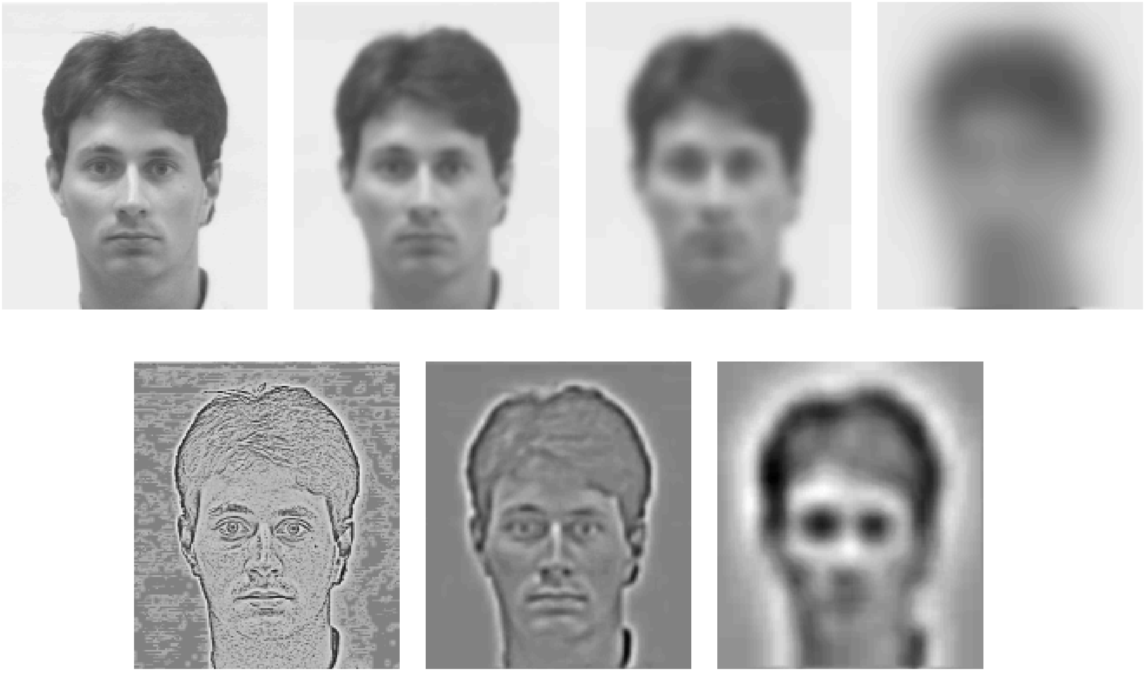

Figure 2: A multi-scale decomposition of an image of the face of a well-known computer scientist. On the top, the face is progessively blurred; the three bottom images are the successive differences. Adding them together, plus the blurriest version, gives back the original face. Note how the different scales contain different information: at the finest, the exact location of edges is depicted; in the middle, the local features are seen but the grey level of the skin and hair are equal; at the coarsest, the global shapes are shown.

is an unobserved random variable $\psi : X \longrightarrow X$ and a normalized signal s_0 such that:

$$s \approx s_0 \circ \psi.$$

In other words, the warping puts s in a more standard form s_0. Here s_0 might be a fixed (i.e. non-random) template, or might itself be random although one aspect of its variability has been eliminated by the warping. Note that in the case where X is a vector space, one can describe the warping ψ as a vector of displacements $\psi(\vec{x}) - \vec{x}$ of specific points. But the components of this vector are numbers representing coordinates of points, not values of the signal, i.e. the domain of s, not the range of s as in the previous class of deformations. This is a frequent confusion in the literature. An example from the PhD thesis of P. Hallinan is given in figure 3.

3. *Hierarchical geometric patterns and parsing the signal.* A fundamental fact about real world signals is that their statistics vary radically from point to point, i.e. they do not come from a so-called ‘stationary process’. This non-stationarity is the basic statistical trace of the fact that the world has a discrete as well as a continuous behavior – it is made up of discrete

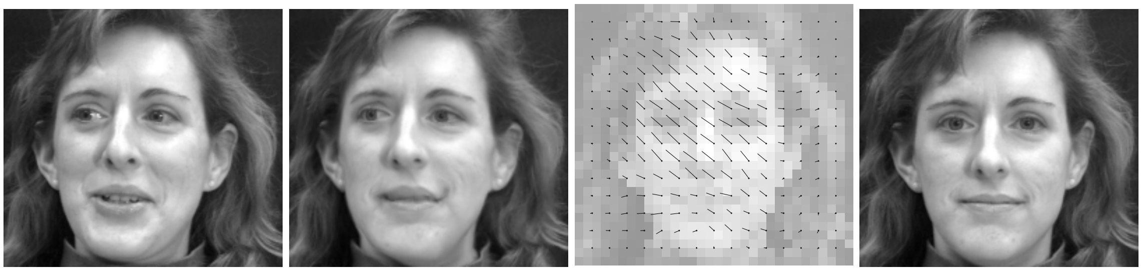

Figure 3: On the far left and far right are two images of the same woman's face. When the warping shown by the arrows in the 3rd image is applied to the face on the right, the second image is produced which nearly matches the one on the left. Note, however, that no teeth are present on the right, so the warping steches the lips to cover the mouth.

events and objects. A central problem in the analysis of such signals is to tease apart these parts of its domain, so as to explicitly label the distinct objects/events affecting this signal. In speech, these distinct parts are the separate phones, phonemes, words, phrases, sentences and speech acts. In vision, they are the various objects, their parts and groupings, present in the viewed scene. Note that in both cases, the objects or processes are usually embedded in each other, hence form a hierarchy. The generic formalism for this is a grammar. Putting in this general setting, if s is a map $X \longrightarrow V$, the basic unobserved random variable is a tree of subsets $X_a \subset X$ typically with labels l_a, such that for every node a with children b:

$$X_a = \bigcup_{a \to b} X_b.$$

Typically, the grammar also gives a stochastic model for an elementary signal $s_t : X_t \longrightarrow R$ for all leaves t of the tree and requires that $s\big|_{X_t} \approx s_t$. The most developed formalism for such grammars are the models called equivalently random branching processes or PCFG's (probabilistic context-free grammars). But most situations require context-sensitive grammars, i.e. the probability of a tree does not factor into terms, one for each node and its children. A parse tree of parts is very natural for the face: the parts correspond to the usual facial features – the eyes, nose, mouth, ears, eyebrows, pupils and eyelids, etc. An example, decomposing a dog, from the work of S.-C. Zhu is shown in figure 4.

What makes the inference of the unobserved random variables in pattern theory hard is not that any of the above models are necessarily hard to use but rather that all of them tend to coexist, and then inference becomes especially hard. In fact, the full model of a signal may involve warping and superposition at many levels and a tree of parse trees may be needed to express the full hierarchy of parts. The world is not simple.

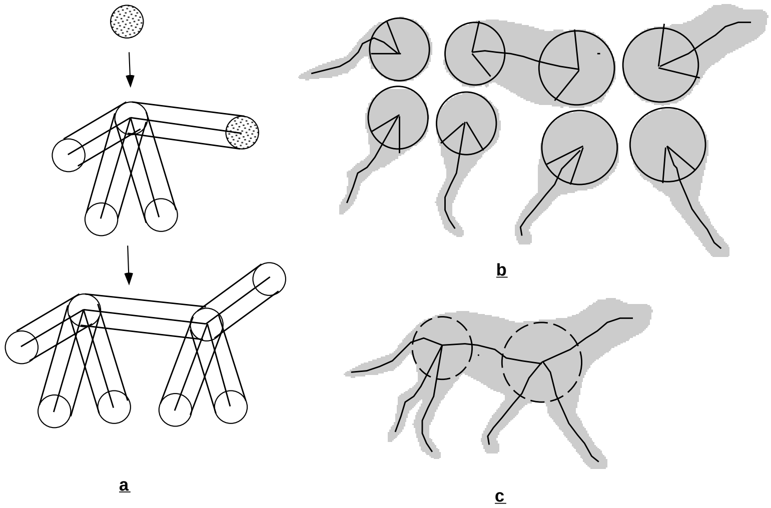

Figure 4: The grammatical decomposition of a dog into parts, from the work of Song-Chun Zhu. On the left, the gramar generates a tree of protrusions for the limbs and a basic strip for the body. On the top right, each part is given a concrete realization as a subset of the plane (with constraints so they match up). On the bottom right, these parts are assembled.

3. Bayesian Probability Theory: Pattern Analysis and Pattern Synthesis

A second element of pattern theory is its use of Bayesian probability theory. An advantage of this Bayesian approach, compared with other vision theories, is that it requires that you first create a stochastic model and, afterwards, you seek algorithms for inferring on the basis of the model. Thus it separates the algorithms from the models. This distinguishes pattern theory from neural network approaches such as multi-layer perceptrons (Hertz91). From our perspective, these theories try to solve two difficult tasks — both modeling and computation — at once. In a Bayesian approach, as in Hidden Markov models and Bayes Nets, we first learn the models and verify them explicitly by stochastic sampling, *and then* seek algorithms for applying the models to practical problems. We believe that learning models and algorithms separately will lead to more tractable problems. Moreover, the explicit nature of the representations leads to a better understanding of the internal workings of an algorithm, and to knowing what problems they will generalize to.

We now give a brief introduction to the techniques of Bayesian probability theory. In general, we wish to infer the state of the random variable S describing the state of the world, given some measurement, some observed random variable I. Thus the variables S would correspond to the variables in our representations of the world, for example the variables representing the shape of a face, while the measurement I would correspond to the observed images. Within the Bayesian framework, one infers S by considering $P(S \mid I)$, the *a posteriori* probability of the state of world given the measurement. Note that by definition of conditional probabilities, we have

$$P(S \mid I)P(I) = P(S , I) = P(I \mid S)P(S).$$

Dividing by $P(\ \boldsymbol{I}\ )$, we obtain Bayes' theorem

$$P(\ S \mid \boldsymbol{I}\ ) = \frac{P(\ \boldsymbol{I} \mid S\ )P(\ S\ )}{P(\ \boldsymbol{I}\ )} = \frac{P(\ \boldsymbol{I} \mid S\ )P(\ S\ )}{\sum_S P(\ \boldsymbol{I} \mid S\ )P(\ S\ )}. \tag{1}$$

This simple theorem re-expresses $P(\ S \mid \boldsymbol{I}\ )$, the probability of the state given the measurement, in terms of $P(\ \boldsymbol{I} \mid S\ )$, the probability of observing measurement given the state, and $P(\ S\ )$, the probability of the state. Each of the terms on the right-handside (RHS) of the above equation has an intuitive interpretation.

The expression $P(\ \boldsymbol{I} \mid S\ )$, often termed the *likelihood function*, is a measure of how likely a measurement is given we know the state of the world. In this book, $\boldsymbol{I}$ is usually an image and this function is also called the *imaging model*. To see this, note that given we know the state of the world, e.g. the light sources, the objects, and the reflectance properties of the surfaces of the objects, then we can re-create, as an image, our particular view of the world. Yet, due to noise in our imaging system and imprecision of our models, this re-creation will have an implicit degree of variability. Thus, $P(\ \boldsymbol{I} \mid S\ )$ probabilistically models this variability.

The expression $P(\ S\ )$, referred to as the *prior model*, models our prior knowledge about the world. In vision, one often says that the prior is necessary because the reconstruction of the 3D world from a 2D view is not "well-posed." A striking illustration is given by "Mooney faces": images of faces in strong light which illuminates part of the face to saturation and leaves the rest black. These images tend to be confusing at first and then to suddenly look like a quite realistic face. But if these images are presented with the opposite contrast, they are nearly unrecogizable. We interpret this to mean that we have a lot of information on the appearance of faces in our learned prior and this information can be used to fill in the missing parts of the face. But if the contrast is reversed, we cannot relate our priors to this image at all.

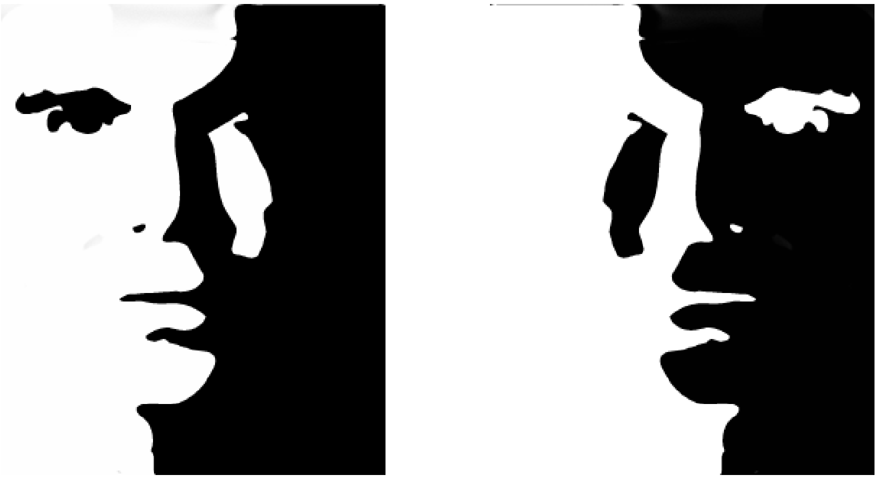

Figure 5: On the left, a 'Mooney' face of a well-known psychophysicist. On the right, the same image with contrast and left/right reversed, making it nearly unidentifiable.

As with the Mooney face example, in general the image alone is not sufficient to determine the scene and, consequently, the choice of priors becomes critically important. They embody the knowledge of the patterns of the world that the visual system uses to make valid 3D inferences. Some such assumptions have been proposed by workers in biological vision and include Gibson's *ecological constraints* and Marr's *natural constraints*. But more than such general principles, we need probability models on representations which are sufficiently rich to model all the important patterns of the world. It is becoming increasingly clear that fully non-parametric models need to be learned to effectively model virtually all non-trivial classes of patterns (see for example the texture modeling in Chapter 4).

What can we do with a Bayesian model of the patterned signals in the world? On the one hand, we can use it to perform probabilistic inference such as finding the most probable estimate of the state of the world contingent on having observed a particular signal. This is called the MAP or *maximum a posteriori* estimate of the state of the world. On the other hand, we can sample from the model, for example fixing some of the world variables S, and using this distribution to construct sample signals I generated by various classes of objects or events. A good test of whether the prior has captured all the patterns in some class of signals is to see if these samples are good imitations of life. From a pattern theory perspective, the analysis of the patterns in a signal and the synthesis of these signals are inseparable problems and use a common probabilistic model: computer vision should not be separated from computer graphics, nor speech recognition from speech generation.

It is helpful to also consider the patterns in signals from the perspective of information theory (Cover-Thomas 91). This approach has its roots in work of Barlow (61) (see also (Rissanen 89)). The idea is that instead of writing out any particular perceptual signal I in raw form as a table of values, we seek a method of encoding I which minimizes its expected length in bits: i.e. we take advantage of the patterns possessed by most I to encode them in a compressed form. We consider coding schemes which involve choosing various auxiliary variables S and then encoding the particular I using these S(e.g. S might determine a specific typical signal I_S and we then need only to encode the deviation $I - I_S$). We write this:

$$\text{length}(\text{code }(I, S)) = \text{length}(\text{code }(S)) + \text{length}(\text{code }(I \text{ using } S)). \tag{2}$$

The mathematical problem, in the information theoretic setup, is, for a given I, to find the S leading to the shortest encoding of I, and moreover, to find the encoding *scheme* leading to the shortest expected coding of all I's. This optimal choice of S is called the *minimum description length* or MDL estimate of S:

$$\text{MDL est. of } S = \arg\min_S [\ \text{length}(\text{code }(S)) + \text{length}(\text{code }(I \text{ using } S))]. \tag{3}$$

There is a close link between the Bayesian and the information-theoretic approaches which comes from Shannon's optimal coding theorem. This theorem states that given a class of signals I, the coding scheme for such signals for which a random signal has the smallest expected length satisfies:

$$\text{length}(\text{code }(\ I\ )) = -\log_2 P(\ I\ ) \tag{4}$$

(where fractional bit lengths are achieved by actually coding several I's at once, and doing this the LHS gets asymptotically close to the RHS when longer and longer sequences of signals are encoded at once). We may apply Shannon's theorem both to encoding S and to encoding I, given S. For these encodings len(code(S)) $= -\log_2 p(\ S\ )$ and len(code(I using S)) $= -\log_2 p(\mathbf{I}|\ S\ )$. Therefore, taking $\log_2$ of equation (1), we get equation (4) and find that the most probable estimate of S is the same as the MDL estimate.

Finally, pattern theory also suggests a general framework for algorithms. 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 partially 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 signals with parsed structure where one component of the signal partially obscures another. When this happens, 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.

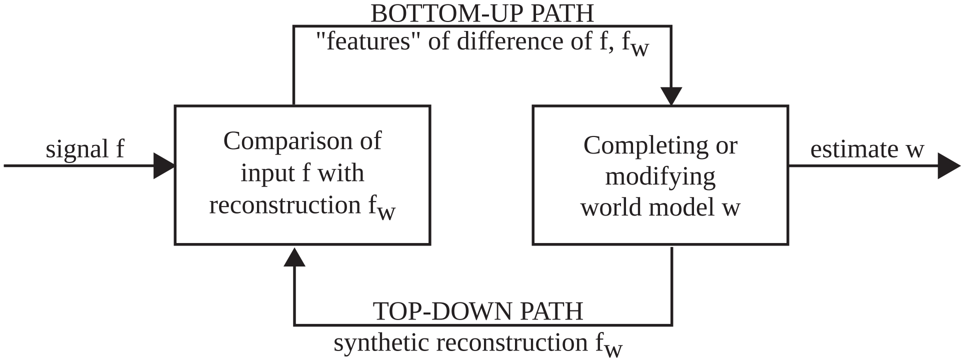
BOTTOM-UP PATH
"features" of difference of f, f_W
signal f
Comparison of input f with reconstruction f_W
Completing or modifying world model w
estimate w
TOP-DOWN PATH
synthetic reconstruction f_W

Figure 6: The fundamental architecture of pattern theory

This framework uses signal synthesis in an essential way and this requirement for feedback gives an intriguing relation to the known properties of mammalian cortical architecture (Mumford 91). Note that, although stemming from similar ideas, the idea of 'analysis by synthesis' is logically separate from the Bayesian formulation and other aspects of pattern theory. Thus the Bayesian approach might be carried out with other algorithms and 'analysis by synthesis' might be used to implement other theories.

A variant of this architecture has been introduced in tracking algorithms (Isard-Blake 96) and is applicable whenever sources of information become available or are introduced in stages. Instead of seeking the most probable values of the world variables S in one step, suppose you *sample* from the posterior distribution $P(\ S \mid \boldsymbol{I}\ )$. The aim to sample sufficiently well so that no matter what later signal $\boldsymbol{I'}$ or later information on the world S' arrives, you can calculate an updated sample of $P(\ S \mid \boldsymbol{I}\ , I', S')$ from the earlier sample.

A final word of caution: at this point in time, no algorithms have been devised and implemented which can duplicate in computers human abilities to perceive the patterns in the signals of the five senses. Pattern theory is a beautiful theoretical analysis of the problems but very much a work in progress when it comes to the challenge of creating a robot with human-like perceptual skills.