MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Technical Report No. 1548 June, 1995

Alignment by Maximization of Mutual Information

Paul A. Viola

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

A new information-theoretic approach is presented for finding the pose of an object in an image. The technique does not require information about the surface properties of the object, besides its shape, and is robust with respect to variations of illumination. In our derivation, few assumptions are made about the nature of the imaging process. As a result the algorithms are quite general and can foreseeably be used in a wide variety of imaging situations.

Experiments are presented that demonstrate the approach registering magnetic resonance (MR) images with computed tomography (CT) images, aligning a complex 3D object model to real scenes including clutter and occlusion, tracking a human head in a video sequence and aligning a view-based 2D object model to real images.

The method is based on a formulation of the mutual information between the model and the image called EMMA. As applied here the technique is intensity-based, rather than feature-based. It works well in domains where edge or gradient-magnitude based methods have difficulty, yet it is more robust than traditional correlation. Additionally, it has an efficient implementation that is based on stochastic approximation.

Finally, we will describe a number of additional real-world applications that can be solved efficiently and reliably using EMMA. EMMA can be used in machine learning to find maximally informative projections of high-dimensional data. EMMA can also be used to detect and correct corruption in magnetic resonance images (MRI).

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract #:N00014-94-01-0994. Paul Viola was also supported by USAF ASSERT program, Parent Grant#:F49620-93-1-0263.

Alignment by Maximization of Mutual Information

by

Paul A. Viola

Submitted to the Department of Electrical Engineering and Computer Science on June 1995, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

Over the last 30 years the problems of image registration and recognition have proven more difficult than even the most pessimistic might have predicted. Progress has been hampered by the sheer complexity of the relationship between an object and its image, which involves the object's shape, surface properties, position, and illumination.

Changes in illumination can radically alter the intensity and shading of an image. Nevertheless, the human visual system can use shading both for recognition and image interpretation. We will present a measure for comparing objects and images that uses shading information, yet is explicitly insensitive to changes in illumination. This measure is unique in that it compares 3D object models directly to raw images. No pre-processing or edge detection is required. We will show that when the *mutual information* between model and image is large they are likely to be aligned. Toward making this technique a reality we have defined a concrete and efficient technique for evaluating entropy called EMMA.

In our derivation of mutual information based alignment few assumptions are made about the nature of the imaging process. As a result the algorithms are quite general and can be used in a wide variety of imaging situations. Experiments demonstrate this approach aligning a number of complex 3D object models to real images. In addition, we demonstrate that the same technique can be used to solve problems in medical registration.

Alignment is accomplished by adjusting the pose of an object until the mutual information between image and object is maximized. We will present a gradient descent alignment procedure based on stochastic approximation that has a very efficient implementation. For this application stochastic approximation affords a speed up of at least a factor of 500 over gradient descent. In addition, stochastic approximation can be used to accelerate a variety of other vision applications. We will describe an existing vision application which can be accelerated by a factor of 30 using stochastic approximation.

Finally, we will describe a number of additional real-world applications that can be solved efficiently and reliably using EMMA. EMMA can be used in machine learning to find maximally informative projections of high-dimensional data. EMMA can also be used to detect and correct corruption in magnetic resonance images (MRI).

Thesis Committee:	Prof. Tomas Lozano-Perez (Co-Supervisor)
	Prof. Christopher G. Atkeson (Co-Supervisor)
	Prof. W. Eric L. Grimson
	Prof. Berthold K. P. Horn

Acknowledgments

I would like to acknowledge the MIT Artificial Intelligence Laboratory for being a haven of intellectual freedom. Professors Tomás Lozano-Pérez, Christopher Atkeson, Eric Grimson, Rodney Brooks, and others have worked to build an environment where every resource is unfettered. Chris and Tomás have supported me through thick and through thin. They've taught me how truly valuable unconditional support can be.

Students are the AI lab's most precious resource and I benefited from discussions with Phil Agre, Davi Geiger, David Chapman, Jose Robles, Tao Alter, Misha Bolotski, Jonathan Connel, Karen Sarachik, Maja Mataric, Ian Horswill, Colin Angle, Cynthia Ferrel, Henry Minsky, Saed Younis, Rick Lathrope, Barbara Moore, and many others. Willliam Wells has stood out among all those I've met at MIT. His unique approach to vision, his care in research and his advice as a friend have proven invaluable. I owe much to John Shewchuk, then of Brown University, for introducing me to the field of statistical learning. Memories of his irrepressible intellectual curiosity serve to continually motivate my own thinking.

Outside of MIT I have spent many productive months at the Salk Institute in the Computational Neurobiology Laboratory of Terrence Sejnowski. Terry is the most tirelessly devoted scientist that I have ever met. In his lab I learned that science is an uncompromising pursuit of truth. Science is about building on the work of others, and that to build one must first understand. In Terry's lab I have had the pleasure of working with David Lawrence, Rich Zemel, Nici Shraudolph, Tony Bell, and Peter Dayan. Each of them has had an effect on some part of this thesis.

Most importantly I must recognize the technical contributions of Sara Billey. She worked tirelessly with me to understand the most cryptic mathematics and clear up my own cryptic thinking. Without her this thesis would not exist.

To:

Sara Billey for being the love of my life.

My parents Mary Ancona-Viola and Alfredo Viola for making it all possible.

Contents

Chapter 1

Introduction

This thesis is about a new information theoretic approach for solving several standing problems in computer vision and image processing. For example, this approach can be used to find the correct alignment between a three dimensional model and an image. While alignment is a critical component of the object recognition problem, it is also useful by itself in medical and military applications. We will also describe several other applications, including an image processing application and a new form of unsupervised learning. While the form of these applications is quite different the underlying theory and derivations are very similar. Preliminary investigation imply that the theory presented here will have wide application.

Computer vision has proven more difficult than even the most pessimistic might have predicted. While the problem has been of interest for over 30 years, progress has been painstakingly slow. Even the best computer vision systems stand in stark contrast to the human visual system: our perception of images is effortless and robust; computer vision systems are at best slow and unreliable. Among other difficulties, progress has been hampered by the sheer complexity of the relationship between image and object, which involves the object's shape, surface properties, position, and illumination.

A computer vision program is faced with the task of interpreting an image of intensities. While information about the shape and location of objects is somehow embedded in these intensities, the actual intensities that arise in an image are difficult to interpret. For example, changes in illumination can radically alter the intensity and shading of an image. Though the human visual system can use shading both for recognition and image interpretation, most existing computer object recognition systems cannot. These systems throw out shading information in an effort to obtain "illumination invariance". We will present a measure for comparing objects and images that uses shading information, yet is explicitly insensitive to changes in illumination. This measure is unique in that it compares 3D object models directly to raw images. No pre-processing or edge detection is required.

This image/model comparison measure has been rigorously derived from information theory. Both the theory and algorithms involved are new, and are based on a efficient scheme for evaluating mutual information called EMMA[1]. The derivation of the the alignment procedure requires few assumptions about the nature of the imaging process. As a result the algorithms are quite general and can be used in a wide variety of imaging situations. Experiments demonstrate that this approach can align a number of complex 3D object models to real images. In addition, the same technique can be used to solve problems in medical registration.

Alignment adjusts the pose of an object until the mutual information between image and object is maximized. Pose adjustment can be accomplished by ascending the gradient of mutual information. We will present an alignment procedure based on stochastic approximation that affords a speed up of at least a factor of 500 over gradient ascent. In addition, stochastic approximation can be used to accelerate a variety of other vision applications. We will describe an existing vision application which can be accelerated by a factor of 30 using stochastic approximation.

EMMA has also proven useful in a number of tasks beside alignment. For example, an entropy minimization framework that can be used to detect and correct corruption in magnetic resonance images (MRI). EMMA can also be used to define a new form of *unsupervised learning*. Unsupervised learning has been popularized in the neural network literature as a scheme for simplifying the representations of complex data. EMMA can be used to find low-dimensional projections of a high dimensional input space that are maximally informative.

[1] EMMA is a random but pronounceable subset of the letters in the words "EMpirical entropy Manipulation and Analysis".

1.1 An Introduction to Alignment

The general problem of alignment entails comparing a predicted image of an object with an actual image. Given an object model and a pose (coordinate transformation), a model for the imaging process could be used to predict the image that will result. If we had a good imaging model then deciding whether an image contained a particular model at a given pose is straightforward: compute the predicted image and compare it to the actual image directly. Given a perfect imaging model the two images will be identical, or close to it. Of course finding the correct alignment is still a remaining challenge.

The relationship between an object model (no matter how accurate) and the object's image is a complex one. The appearance of a small patch of a surface is a function of the surface properties, the patch's orientation, the position of the lights and the position of the observer. Given a model $u(x)$ and an image $v(y)$ we can formulate an imaging equation,

$$v(T(x)) = F(u(x), q) \tag{1.1}$$

or equivalently,

$$v(y) = F(u(T^{-1}(y)), q) \ . \tag{1.2}$$

The imaging equation is separable into two distinct components. The first component is called a transformation, or pose, denoted T. It relates the coordinate frame of the model to the coordinate frame of the image. The transformation tells us which point in the model is responsible for a particular point in the image. The second component is the imaging function, $F(u(x), q)$. The imaging function determines the value of image point $v(T(x))$. In general a pixel's value may be a function both of the model and other exogenous factors. For example an image of a three dimensional object depends not only on the object but also on the lighting. The parameter, q, collects all of the exogenous influences into a single vector.

One reason that it is, in principle, possible to define F is that the image does convey information about the model. Clearly if there were no mutual information between u and v, there could be no meaningful F. We propose to finesse the problem of finding and computing F by dealing with this mutual information directly. We will present an algorithm that aligns by maximizing the mutual information between model and image. It requires no a priori model of the relationship between surface properties and scene intensities – it only assumes that the model tells more about the scene when it is correctly aligned.

1.1.1 An Alignment Example

One of the alignment problems that we will address involves finding the pose of a three-dimensional object that appears in a video image. This problem involves comparing two very different kinds of representations: a three-dimensional model of the shape of the object and a video image of that object. For example, Figure 1.1 contains a video image of an example object on the left and a depth map of that same object on the right (the object in question is a person's head: RON). A depth map is an image that displays the depth from the camera to every visible point on the object model. A depth map is a complete description of the shape of the object, at least the visible parts.

From the depth map alone it might be difficult to see that the image and the model are aligned. The task can be made much easier, at least for us, if we simulate the imaging process and construct an image from the 3D model. Figure 1.2 contains two computer graphics renderings of the object model. These synthetic images are constructed assuming that the 3D model has a Lambertian surface and that the lighting comes from the right. It is almost immediately obvious that the model on the left is more closely aligned to the true image than the model on the right. Unfortunately, what we find trivial is very difficult for a computer. The intensities of the true video image and the synthetic images are very different. The true image and the correct model image are in fact uncorrelated. Yet any person can glance at these images and decide that both are images of a head and that both heads are looking in roughly the same direction. The human visual system is capable of ignoring the superficial differences that arise from changes in illumination and surface properties.

It is not easy to build an automated alignment procedure that can make this kind of comparison. It is harder still to construct a system that can find the correct model pose. We have built such a system. That system selected the pose of the model shown at left in Figure 1.2.

As mentioned above, the synthetic images of RON were generated under the assumption the model surface is Lambertian and the lighting is from the right. Lambert's law is perhaps the simplest model of surface reflectivity. It is an accurate model of the reflectance of a matte

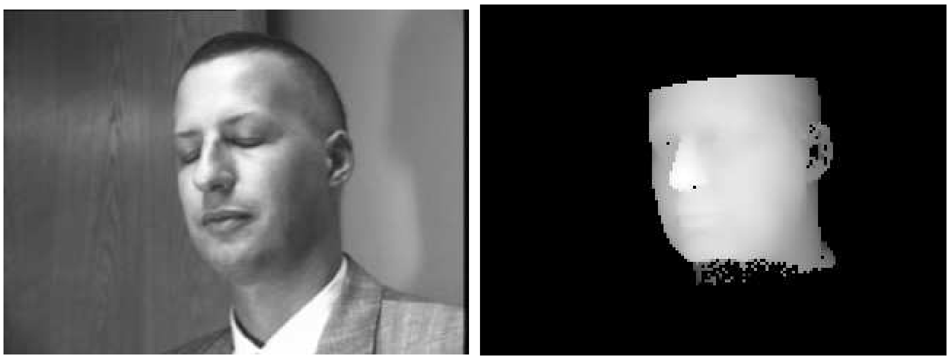


Figure 1.1: Two different views of RON. On the left is a video image. On the right is a depth map of a model of RON. A depth map describes the distance to each of the visible points of the model. White denotes points that are closer, black further.

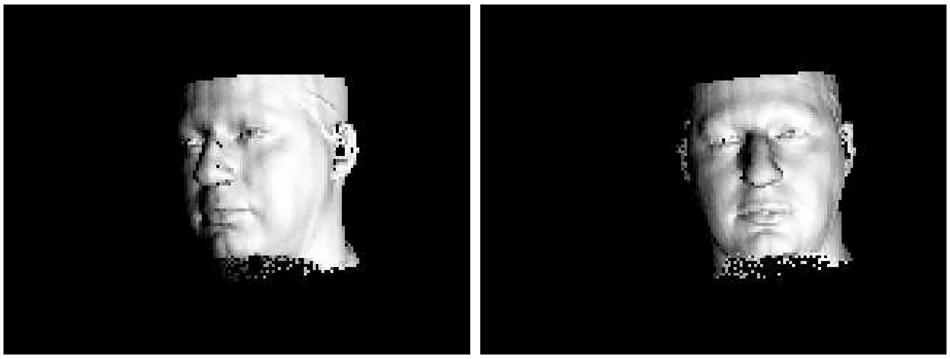


Figure 1.2: At left is a computer graphics rendering of a 3D model of RON. The position of the model is the same as the position of the actual head. At right is a rendering of the head model in an incorrect pose.

or non-shiny surface. Lambert's law states that the visible intensity of a surface patch is related to the dot product between the surface normal and the lighting. For a Lambertian object the imaging equation is:

$$v(T(x)) = \sum_i \alpha_i \vec{l_i} \cdot u(x) \ , \tag{1.3}$$

where the model value $u(x)$ is the normal vector of a surface patch on the object, l_i is a vector pointing toward light source i, and α_i is proportional to the intensity of that light source ((Horn, 1986) contains an excellent review of imaging and its relationship to vision).

Drawing the explicit parallel between (1.1) and (1.3) we can see that the imaging function is

$$F(u(x), q) = \sum_i \alpha_i \vec{l_i} \cdot u(x) \quad , \tag{1.4}$$

where $q = \{\alpha_i, \vec{l_i}\}$. As the illumination changes the functional relationship between the model and image will change.

Since we can not know beforehand what the imaging function will be, aligning a model and image can be quite difficult. These difficulties are only compounded if the surface properties of the object are not well understood. For example, many objects can not be modeled as having a Lambertian surface. Different surface finishes will have different reflectance functions. In general reflectance is a function of lighting direction, surface normal and viewing direction. The intensity of an observed patch is then:

$$v(T(x)) = \sum_i R(\alpha_i, \vec{l_i}, \vec{o}, u(x)) \quad , \tag{1.5}$$

where $\vec{o}$ is a vector pointing toward the observer from the patch and $R(\cdot)$ is the reflectance function of the surface. For an unknown material a great deal of experimentation is necessary to completely categorize the reflectance function. Since a general vision system should work with a variety of objects and under general illumination conditions, overly constraining assumptions about reflectance or illumination should be avoided.

Let us examine the relationship between a real image and model. This will allow us to build intuition about the alignment process. Data from the real reflectance function can be obtained by aligning a model to a real image. An alignment associates points from the image with points from the model. If the alignment is correct, each pixel of the image can be interpreted as a sample of the imaging function $R(\cdot)$. The imaging function could be displayed by plotting intensity against lighting direction, viewing direction and surface normal. Unfortunately, because intensity is a function of so many different parameters the resulting plot can be prohibitively complex and impossible to visualize. Significant simplification will be necessary if we are to detect any structure in this data.

In a wide variety of real images we can assume that the light sources are far from the object (at least in terms of the dimensions of the object). When this is true and there are no shadows, each patch of the object will be illuminated in the same way. Furthermore, we will assume that the observer is far from the object, and that the viewing direction is therefore constant throughout the image. The resulting relationship between normal and intensity is three dimensional: the normal vector has unit length and is determined by two parameters, its x and y components; the intensity is a third parameter. A three dimensional scatter plot of normal versus intensity is really a slice through the high dimensional space in which $R(\cdot)$ is defined. Though this graph is much simpler than the original, three dimensional plots are still quite difficult to interpret. We will slice the data once again so that all of the points have a single value for the y component of the normal.

Figure 1.3 contains a graph of the intensities along a single scan-line of the image of RON. Figure 1.4 shows similar data for the correctly aligned model of RON. Model normals from this scan-line are displayed in two graphs: the first shows the x component of the normal while the second shows the y component. Notice that we have chosen this portion of the model so that the y component of the normal is almost constant. As a result the relationship between normal and intensity can be visualized in only two dimensions. Figure 1.5 shows the intensities in the image plotted against the x component of the normal in the model. Notice that this relationship appears both consistent and functional. Points from the model with similar surface normals have very similar intensities. The data in this graph could be well approximated by a smooth curve. We will call an imaging function like this one *consistent*. Interestingly, we did not need any information about the illumination or surface properties of the object to determine that there is a consistent relationship between model normal and image intensity.

Figure 1.6 shows the relationship between normal and intensity when the model and image are no longer aligned. The only difference between this graph and the first is that the intensities come from a scan-line 3 centimeters below the correct alignment (i.e. the model is no longer aligned with the image, it is 3 centimeters too low). The normals used are the same. The resulting graph is no longer *consistent*. It does not look as though a simple smooth curve would fit this data well.

In summary, when model and image are aligned there will be a consistent relationship between image intensity and model normal. This is predicted by our assumption that there is an imaging function that relates models and images. While the actual form of this function depends on lighting and surface properties, a correct alignment will always lead to a consistent relationship. Conversely, when model and image are misaligned the relationship between

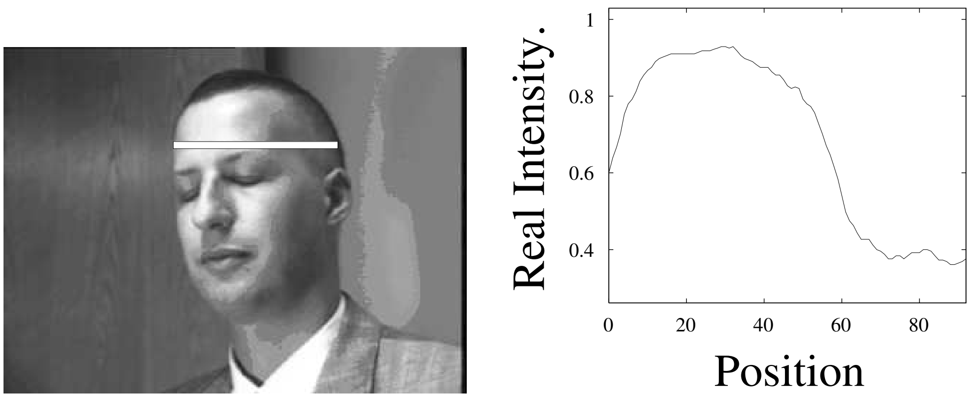


Figure 1.3: On the left is a video image of RON with the single scan-line highlighted. On the right is a graph of the intensities observed along this scan line.

intensity and normal is inconsistent.

A major contribution of this thesis is the derivation of a formal technique that delivers a principled estimate of "consistency". We will show that when the mutual information between an image and a model is high they are likely to be aligned. Toward making this technique a reality we have defined a new approach for evaluating entropy and information called EMMA. We have also defined an efficient scheme for adjusting a set of parameters so that mutual information and entropy can be optimized. We will use EMMA to effectively evaluate and adjust the alignment of three dimensional models and two dimensional images.

This same technology can be used for the alignment of other types of signals. In its full generality, EMMA can be used whenever there is a need to align images from two different sensors, the so-called "sensor fusion" problem. For example, in medical imaging data from one type of sensor (such as magnetic resonance imaging) must be aligned to data from another sensor (such as computed tomography). We will demonstrate that EMMA can be used to solve problems such as this.

Though developed for alignment, the EMMA estimates of entropy and mutual information can be used in other applications. We will show that EMMA can be used to correct inhomogeneities in MRI scans. In addition, we will derive a new approach for dimensionality reduction based on entropy. Similar to principal components analysis, our technique can find low dimensional projections of higher dimensional data that preserve the most information

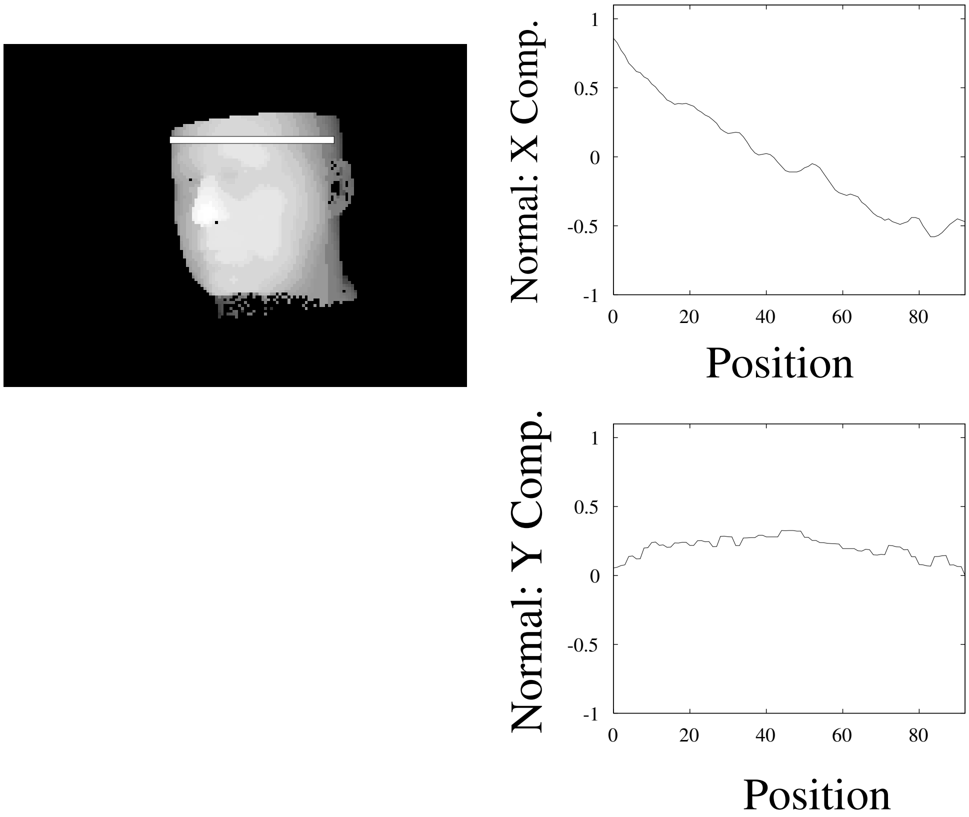


Figure 1.4: On the left is a depth map of RON with the single scan-line highlighted. At top right is a graph of the x component of the surface normal. On the bottom right is the y component of the normal.

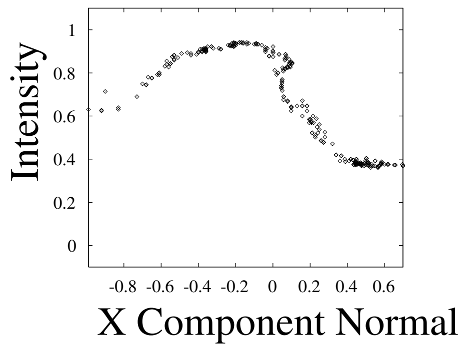
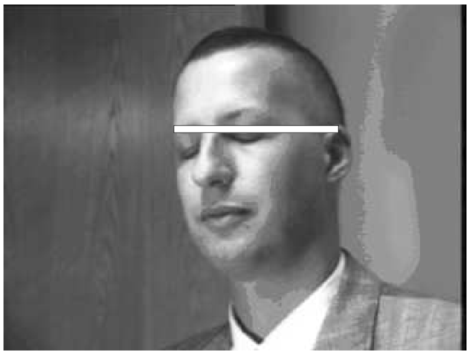


Figure 1.5: The Aligned Case: A scatter plot of the intensity of the video image versus the x component of the surface normal from the model. The image and model are correctly aligned.



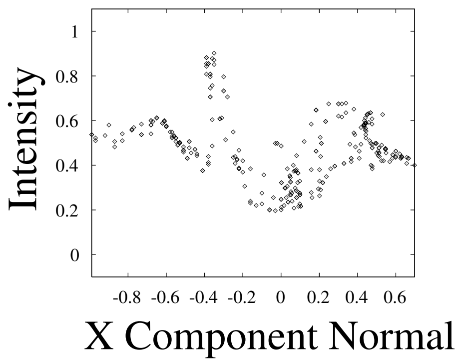


Figure 1.6: The Misaligned Case: On the left is the misaligned scan-line from the video image of Ron. On the right is a scatter plot of the intensity of this part of the video image versus the x component of the surface normal from the model.

possible.

1.2 Overview of the Thesis

The second chapter contains an overview of the probability theory necessary to understand what EMMA is doing and how it does it. The third chapter discusses estimation of entropy from samples. While a number of techniques for manipulating entropy currently exist, EMMA combines computational efficiency with the flexibility necessary to model a wide variety of distributions. The fourth chapter returns to our discussion of alignment. We show here that EMMA is capable of aligning signals where simpler techniques cannot. This chapter will present the basic equations that underly alignment by maximization of mutual information. The fifth chapter contains a wide variety of alignment experiments designed both to validate our approach and explore the scope of possible application. In chapter six we will describe applications besides alignment to which we have applied EMMA. For example, our scheme for efficiently manipulating entropy includes a stochastic form of gradient descent. We will describe a flow estimation problem in which stochastic gradient descent speeds convergence by a factor of thirty. Chapter seven will include a discussion of our results and a comparison with related work.

Chapter 2

Probability and Entropy

One of the key insights in this thesis is that many of the techniques that are common in computer vision, such as correlation, are easily interpreted as statistics of random variables. Once this is done we can use a broad range of tools from probability to analyze the behavior of these statistics. The theory of probability will help us determine how statistics converge, what they converge to, and more importantly how alternative statistics might be more appropriate.

In this chapter we will introduce the basic mathematics that underly probability. In subsequent chapters we will assume that the reader has a fairly thorough knowledge of probability, statistics, entropy, and coding. This chapter is intended both as a review of the required techniques and theorems and as a bridge for a reader unfamiliar with these topics. The final sections of this chapter contain a new analysis and discussion of Parzen density estimation. Parzen density estimation will play an important role in the estimation of entropy in subsequent chapters.

We will sometimes use a simplified, or looser, definition of concepts like events and random variables than is typical. If you get overly confused reading this chapter, any good book on probability should clear things up (Papoulis, 1991; Baclawski et al., 1990). In general we will leave out the proofs of anything that is easily looked up, and of course most of the theory presented is cited here without reference. Unfortunately probability and statistics seems to have many conflicting standard notations. In our own definitions we will try to be consistent with the prevailing conventions.

2.1 Random Variables

In many cases an algebraic model of a physical system allows us to accurately predict its behavior. For instance, circuit theory can be used to analyze a particular circuit and predict that when a switch is closed current will flow. The physics of many circuits can be modeled as equations where unknown quantities are recorded as variables. In the case of a switched circuit, we can model the resistance of a switch as a variable that can take on one of two values: zero when closed or infinity when open. The current that flows through that resistor can then be predicted from algebraic manipulations. Conversely, knowing the value of the current allows us to predict whether the switch is open or closed. The equivalence of a circuit and a circuit model is fundamental within the fields of physics and engineering.

In a wide variety of physical systems the behavior of particular measurements cannot be easily predicted. The voltage of a wire may be a complex function of the circuit and the thermal noise in a resistor. Even when all of the other circuit variables are known, the voltage cannot be predicted accurately. Luckily, all hope is not lost. We may not know the actual voltage but we may know that it will be "near" V_0, and that it is never, in our experience, higher than V_{max}. Probability, random processes, and random variables provide the tools to quantify the intuitive concepts of "near" and "never".

A random variable, or RV, is a variable whose value is unpredictable. Recall that a variable is a symbol X and a set of values Ω_X over which the variable can range. For example, X could range over the real numbers between 1 and 10. In this thesis we will assume Ω_X will always be a subset of the real numbers. A random variable X is a variable together with a function $P_X : \Omega_X \rightarrow [0, 1]$ called a probability distribution. For example we can construct an RV that models the roll of a six sided die. If the die is "fair" we cannot in advance know what its value will be, but we do know that its value will be one of 6 integers from 1 to 6, and that each will appear roughly one sixth of the time. The RV that describes this die includes the variable symbol X, a sample space $\Omega_X = \{1, 2, 3, 4, 5, 6\}$ of possible outcomes, and a probability distribution function $P_X(n)$ which tells us the probability that X will take on the value n. A particular value of an RV is called a *trial*, for example from a die roll. A collection of trials is called a *sample*. An *event* is a set A such that $A \subset \Omega_X$. The probability of an event, $P_X(X \in A)$ is the proportion of times that you expect to see event A in a large sample. The sum over the sample space of the probability distribution equals one:

$$\sum_{x_i \in \Omega_X} P(X \in \{x_i\}) = 1 \quad .$$

Here we denote the elements of the sample space Ω_X with the lower case letter x. In many cases we will write $P_X(x_i)$, $P(X = x_i)$ or $P(x_i)$ for $P_X(X \in \{x_i\})$[1]. An RV which takes on a finite or discrete set of values is known as a *discrete random variable.* An RV whose range includes some infinite set of continuous values is known as a *continuous random variable.*

A bit of thought leads one to a conundrum regarding continuous RV's—since there are an infinite number of possible outcomes the probability of almost every outcome will be zero. This will in fact be a continuing annoyance to us as we move toward the definition of entropy. Instead of probability distributions for continuous RV's we use probability densities:

$$p_X(x_0) = \lim_{\delta \to 0} \frac{P(x_0 < X < x_0 + \delta)}{\delta} \quad .$$

The probability of an event can just as easily be defined from the density by

$$P(x_{low} < X < x_{high}) = \int_{x_{low}}^{x_{high}} p_X(x) dx \quad .$$

The probability density of an RV always integrates to 1,

$$\int_{-\infty}^{\infty} p_X(x) dx = P_X(-\infty < X < \infty) = 1 \quad .$$

It is not true, however, that $0 \leq p_X(x) \leq 1$. Probability densities are always non-negative, but can have arbitrarily large values. Often densities can be manipulated in the same way that distributions are. In subsequent discussion we will avoid duplication whenever definitions and theorems are the same for both distributions and densities.

[1] Typically probability books say that this is only done when there is no chance for confusion. We know better.

Simple Statistics

A random variable model of a process allows us to answer a variety of quantitative questions about the behavior of the process. Though the voltage across a resistor is unpredictable, its long term average is not. Let us define the intuitive notion of "long term average" as the *expected value* or *mean* of an RV. The expected value $E_X[X]$ is defined as:

$$E_X[X] \equiv \sum_{x_i \in \Omega_X} x_i P(X = x_i) \ , \tag{2.1}$$

or

$$E_X[X] \equiv \int x p(X = x) dx \ . \tag{2.2}$$

For notational convenience we will sometimes refer to the expectation of X as $E[X]$. The mean of a random variable is a deterministic function of its distribution. Intuitively $E[X]$ is the average of the RV's value over a large sample. We will denote a sample a, somewhat non-standardly, by an ordered collection of trials x_a,

$$a = [...x_a...] \ .$$

The size of a sample $\|a\|$ we will refer to as N_a. In a small abuse of notation we will write

$$E_a[X] \equiv \frac{1}{N_a} \sum x_a \ ,$$

for the average over the sample a. Unlike the mean, the sample mean is a random variable. The law of large numbers allows us to prove that in the limit the sample mean equals the expectation:

$$E_X[X] = \lim_{N_a \to \infty} E_a[X] = \lim_{N_a \to \infty} \frac{1}{N_a} \sum x_a \ . \tag{2.3}$$

The mean is an example of a *statistic*. Statistics are deterministic values computed from an RV that sum up its gross or long term behavior. Statistics of X are defined as the expectation of functions of X or possibly $P(X)$.

By itself, $E[X]$ does not tell us much about X. For example, the average lottery number does not help us guess what the next lottery number will be. In addition to knowing the mean, we would like to know on average how close samples of X will be to the mean. We can tell that the average lottery number is a useless statistic by the fact that the variation in lottery numbers is huge. One measure of expected variation is called *variance* and is defined as:

$$Var(x) \equiv E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2 \quad . \tag{2.4}$$

The square root of variance is the *standard deviation*, $\sigma(X)$. The standard deviation is a measure of how far, on average, the samples of X will be from $E[X]$.

Though the expectation of an RV is equal to the infinite mean, as in (2.3), we have not explored its relation to the sample mean. Is the sample mean a good estimate for the true mean of an RV? In a qualified sense the answer is yes. The expectation of the sample mean is the same as the expectation:

$$E[E_a[X]] = E[\frac{1}{N_a}\sum x_a] = \frac{1}{N_a}\sum E[x_a] = E[X] \quad .$$

Expectation, because it is defined as an integral, is linear and can be moved inside the summation. The sample mean is often called an *unbiased* estimator of the true mean. But, how close on average will the sample mean be to the true mean? Under the assumption that the different trials of X are independent and identically distributed, the standard deviation of the sample mean is

$$\sigma(E_a[X]) = \frac{\sigma(X)}{\sqrt{N_a}} \quad .$$

Therefore, the standard deviation of the sample mean approaches 0 as N_a approaches infinity. We can conclude that the sample mean is an unbiased estimate for the true mean, and that the quality of the estimate is better for larger samples.

The mean and variance are the zeroth and first elements of an infinite class of moment statistics. These statistics can used to classify the behavior of an RV with ever increasing accuracy.

The Algebra of Random Variables

Random variables are useful descriptions of processes that occur in the real world. RV's can be used in algebraic equations just as variables are. The value of an equation that includes an RV is another random process. A new RV, Y, can be defined from X as $Y = F(X)$. For discrete RV's, the probability distribution of Y is easily defined as: $P_Y(F(n)) = P_X(n)$. For continuous RV's it is not quite as simple,

$$p_Y(F(x)) = \frac{p_X(x)}{\left|\frac{dF(x)}{dx}\right|} \quad .$$

Intuitively this equation tells us to scale down the density at points where $\left|\frac{\partial F(x)}{\partial x}\right|$ is large. In these regions F acts to stretch out X. The density $p_X(x)$ gets diluted by this stretching.

With this new theory of random variables, and many identities that can only really be hinted at here, we can begin to analyze systems such as the noisy circuit described above. We can answer questions like, "If there is random noise in the voltage from a power supply, how much variation will there be in the current across a resistor on the other side of the circuit?" In general this kind of analysis starts from a description of the distribution of one RV and derives the distribution of other functionally related RV's in the system.

Joint and Conditional Distributions

When one RV is a function of another RV, as in $Y = F(X)$, Y and X are said to be dependent. Measuring X allows us to predict Y. It is also possible that two RV's are related but not directly predictable from each other. An example is a noisy voltage source that is powering a noisy current source. Actually measuring voltage tells you something about current, but it doesn't tell you everything. There is still unpredictability that arises from the current source itself. Finally, it is possible that two RV's are completely independent. For example, two different rolls of a fair die are considered independent.

Dependency can be formalized by examining the joint distribution of two RV's, $P(X, Y)$. The joint distribution tells us about the co-occurrence of events from the RVs X and Y. It is a complete description of the random behavior of both X and Y. From the joint distribution one can compute the *marginal distributions*:

$$P(X) = \sum_{y \in \Omega_Y} P(X, Y = y)$$

and

$$P(Y) = \sum_{x \in \Omega_X} P(X = x, Y) \ .$$

Two variables are *independent* if

$$P(X, Y) = P(X) \cdot P(Y) \ . \tag{2.5}$$

They are considered dependent when the joint distribution is not the product of marginal distributions. A closely related distribution, the conditional distribution, $P(Y \mid X)$, is the probability of Y if we knew X. It is defined as:

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)} \ .$$

Complete, functional dependence can be determined from conditional probability when it is the case that for all $x \in \Omega_X$ that

$$P(Y = F(x) \mid X = x) = 1 \ .$$

What is known as Bayes' Law can be concluded from the following equation:

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)} \frac{P(X)}{P(X)} = P(Y \mid X) \frac{P(X)}{P(Y)} \ .$$

Bayes' Law inverts conditional probabilities. It is quite useful in situations where one would like to conclude the distribution of X from a measurement of Y, but in principle all that is known is $P(Y \mid X)$.

2.2 Entropy

Entropy is a statistic that summarizes randomness. The definition of a random variable makes no mention of how random the variable is. Is a lottery number more or less random than the roll of a die? Entropy helps us answer this question. As we will see, the more random a variable is the more entropy it will have. Much additional material on entropy can be found in the excellent textbook by Cover and Thomas (Cover and Thomas, 1991).

Entropy in one form or another is a very old concept. Its origins clearly date back to the first work on thermodynamics in the last century. Nonetheless, most of the credit for defining entropy and promoting its use in data analysis and engineering falls to Shannon (Shannon, 1948). The most straightforward definition of entropy is as an expectation:

$$H(X) = -E_X[\log(P(X))] = -\sum_{x_i \in \Omega_X} \log(P(X = x_i))P(X = x_i) \ .$$

where we define $0 \ \log(0) = 0$ here and elsewhere in the thesis. The classical definition of entropy applies only to discrete random variables. We will present the definition of continuous entropy, known as differential entropy, later. Entropy is typically defined in terms of the logarithm base 2. In that case entropy is given in units of bits.

Entropy is Code Length

One way of measuring randomness is to compose the shortest message that describes either one or a number of trials of an RV. A trial of a fair coin takes one bit of information to encode: a 1 for heads and a 0 for tails. There is no more efficient technique for encoding a single trial[2]. This restriction does not apply to a message that describes a sample of many trials. If the coin in question comes up tails only one time in a thousand, there are a number of straightforward schemes for encoding a sample that require less than one bit per trial. For instance one could send the length of the sample a and then the position of the zeros. It would take $\log(N_a)$ bits to send the length and $\log(N_a)$ bits to send the position of each zero. The length of a message describing N_a flips of our coin would on average be

$$E[length(a)] = \log(N_a) + P(X = 0)N_a \log(N_a) \ .$$

In this coding scheme a message describing one thousand trials will on average take about 20 bits. The number of bits it takes to encode a sample is dependent both on the number of events and the distribution of the random variable.

A discussion and comparison of coding schemes could take up quite a lot of space. Luckily, using the Kraft inequality it can be proven that on average one needs to send $H(X)$ bits to communicate a trial of the random variable X. Furthermore, Shannon showed that is possible

[2]Provided the coin doesn't have two heads.

to construct a code that will take at most $H(X)+1$ bits on average. A simple algorithm discovered by Huffman can construct the shortest possible codes for any random variable. Because entropy is a bound on the code length that is required to transmit a trial, entropy is often called *information*.

Conditional Entropy, Joint Entropy, and Mutual Information

The concept of mutual information plays a critical role in this thesis. One of the key problems that we will need to solve is, "How likely is it that the random variable Y is functionally dependent on X?" In Section 2.1 we saw that two RV's were independent if and only if their joint density was the product of their marginal densities (see (2.5)). Entropy will allow us to quantify the extent to which two RV's are dependent.

Quantifying dependence is very much like quantifying randomness. Total dependence implies that a measurement of one RV completely determines the other (i.e. knowledge of X removes any randomness from Y). Independence is just the opposite (i.e. knowledge of X does not help you predict Y). Just as joint and conditional distributions relate the co-occurrences of two RV's, entropy can be used to relate the predictability of two RV's. Conditional entropy and joint entropy are defined as:

$$H(Y \mid X) \equiv E_X[E_Y[\log(P(Y \mid X))]]$$

and

$$H(Y, X) \equiv E_X[E_Y[\log(P(Y, X))]] \ .$$

Conditional entropy is a measure of the randomness of Y given knowledge of X. Note that it is an expectation over the different events of X, so it measures on average just how random Y is given X. $H(Y \mid X = x)$ is the randomness one expects from Y if X takes on a particular value. Random variables are considered independent when

$$H(Y \mid X) = H(Y)$$

or,

$$H(X, Y) = H(X) + H(Y) \ .$$

As Y becomes more dependent on X, $H(Y \mid X)$ gets smaller. However, conditional entropy by itself is not a measure of dependency. A small value for $H(Y|X)$ may not imply dependence, it may only imply that $H(Y)$ is small. The *mutual information*, MI, between two random variables is given by

$$I(X,Y) = H(Y) - H(Y \mid X) \ . \tag{2.6}$$

$I(X,Y)$ is a measure of the reduction in the entropy of Y given X.

A number of simple logarithm equalities can be used to prove relations between conditional and joint entropy. For instance, conditional entropy can be expressed in terms of marginal and joint entropies:

$$H(Y \mid X) = H(X,Y) - H(X) \ .$$

This allows us to provide three equivalent expression for mutual information (and a useful identity):

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y \mid X) && (2.7)\\ &= H(X) + H(Y) - H(X,Y) && (2.8)\\ &= H(X) - H(X \mid Y) && (2.9)\\ &= I(Y,X) \ . && (2.10) \end{aligned}$$

An extremely useful inequality on expectations, known as Jensen's inequality, allows us to prove that for any concave function F that

$$E[F(X))] \leq F(E[X]) \ .$$

A function is concave when its second derivative is negative everywhere. Using the fact that the logarithm function is concave, Jensen's inequality allows us to prove the following useful inequalities:

$$\begin{aligned} H(X) &\geq 0 && (2.11)\\ H(Y) &\geq H(Y \mid X) && (2.12)\\ I(X,Y) &\geq 0 && (2.13) \end{aligned}$$

2.2.1 Differential Entropy

While a number of the main theorems of entropy apply both to continuous and discrete distributions, a number of other theorems change significantly. The continuous version of entropy is called differential entropy, and is defined as:

$$h(X) \equiv -E_X[\log(p(X))] = -\int_{-\infty}^{\infty} p(x)\log(p(x))dx \quad . \tag{2.14}$$

For the most part differential entropies can be manipulated, and obey the same identities, as entropy. In fact all of the equalities and inequalities of the previous section hold except for (2.11). Throughout the thesis, when entropy is mentioned, it is to be understood as the applicable form of entropy. When the difference matters we will be explicit.

The most perplexing difference between entropy and differential entropy is that there is no longer a direct relationship between $h(X)$ and code length. It is possible to construct examples where differential entropy is negative. This is an implication of the fact that $p(X)$ can take on values greater than 1. Code length, however, is never negative.

Differential entropy does not provide an absolute measure of randomness. Discouragingly, it is even the case that a density with a differential entropy of negative infinity may still be unpredictable. Examples of this sort can be constructed by embedding a discrete process into a continuous space. For example one could model the roll of a die as a continuous RV. The density would then be made up of a series of delta functions centered at the points one through six. A delta function, often called a Dirac delta function, can be defined from a box car function,

$$b(x, x_{low}, x_{high}) = \begin{cases} \frac{1}{x_{high}-x_{low}} & \text{if } x_{low} < x < x_{high} \\ 0 & \text{otherwise} \end{cases} \quad . \tag{2.15}$$

The box car function is defined so that it integrates to one. The delta function is a box car function in the limit as it approaches zero width,

$$\delta(x) = \lim_{\Delta \to 0} b(x, 0, \Delta) \quad . \tag{2.16}$$

The delta function, because it is a box car, integrates to one. It can be shown that

$$f(x_0) = \int_{-\infty}^{\infty} \delta(x_0 - x) f(x) dx \tag{2.17}$$

Furthermore, from the definition of convolution,

$$(f * g)(x) \equiv \int_{-\infty}^{\infty} f(x - x')g(x')dx' \tag{2.18}$$

we can see that the delta function is the identity operator,

$$(\delta * f)(x) = \int_{-\infty}^{\infty} \delta(x - x')f(x')dx' = f(x) \ . \tag{2.19}$$

The density of the continuous model of die rolling can be formulated as

$$p(x) = \sum_{i=1}^{6} \frac{1}{6}\delta(x - i) \tag{2.20}$$

which will integrate to 1. Furthermore if we define the probability of an event as

$$P(x) = \lim_{\Delta \to 0} \int_{x-\Delta}^{x+\Delta} p(x) \ , \tag{2.21}$$

then $P(X = 1) = \frac{1}{6}$, as will the probability of the other events. Finally we can show that the entropy of X is negative infinity,

$$h(X) = -\int_{-\infty}^{\infty} p(x') \log(p(x'))dx' = -\sum_{i=1}^{6} \frac{1}{6} \log(\infty) = -\infty \ , \tag{2.22}$$

yet X is pretty clearly random.

Though differential entropy does not provide an absolute measure of randomness or code length, it does provide a relative measure of these properties. A random variable X is less predictable than Y whenever $h(X) > h(Y)$. Similarly an event from X requires more bits on average to encode than an event from Y.

2.3 Samples versus Distributions

A random variable is a mathematical structure that is intended to model the behavior of a physical process. In some cases there are excellent physical reasons to believe that an RV is an accurate model of a process. In many other cases the properties of a random physical process may well be unknown. In these cases we may still wish to use the theory of probability to analyze a system. The first step is to find an accurate RV model of our data.

In order to insure that our probabilistic inferences will be correct, our model must be as accurate as possible. While it is possible to model every coin as a fair coin, we do so at our peril. It makes much more sense to perform a large number of experiments intended to test the hypothesis, "Is the coin fair?" There are two important intuitions behind finding an accurate model of a random process. First and foremost, you want a model that seems to explain the data well. It might not make sense to guess that a coin is fair if after 500 flips heads has come up 400 times. But it is also important that your model be plausible. If after a lifetime of experience you realize that most coins are pretty fair, than perhaps it makes more sense to assume that 400 heads is unusual but not sufficiently unusual to assume that the coin is biased.

2.3.1 Model Selection, Likelihood and Cross Entropy

The field of statistics provides many tools for testing the validity of random models. A lot of this work shares a particular form, called *maximum likelihood* model selection. The goal is to select the most probable model given a large sample of measurements. Maximum likelihood selection proceeds in steps: (1) guess the definition of a random variable that might model the process; (2) evaluate the "goodness" of the model by computing the probability that the data observed could have been generated by the model, (3) after evaluating many models retain the model that makes the data most probable. The probability of a sample a under the RV X is the probability of the co-occurrence of the trials in a,

$$\ell(a) = P_X(a) = P_X(x_1 = x_{a_1}, x_2 = x_{a_2}, ...) \ . \tag{2.23}$$

The probability of a sample is usually called its *likelihood* and is denoted $\ell(a)$.

Justification for maximum likelihood model selection is based on Bayes' law. The likelihood of a sample is really a conditional probability, $P(a \mid X)$. Bayes' law allows us to turn the conditional around and find the most likely model given the sample,

$$P(X \mid a) = P(a \mid X)\frac{P(X)}{P(a)} \ . \tag{2.24}$$

In order to compute the model likelihood one must multiply the sample likelihood by the correcting factor $\frac{P(X)}{P(a)}$. The unconditioned probability of the sample $P(a)$ could well be arbitrary, because the sample is the same for all of the models we will evaluate. The prior probability of the model, $P(X)$, poses more problems. Maximum likelihood model selection assumes that all of the models that are to be evaluated are equally likely to have occured, i.e. $P(X)$ is constant. As a result $\frac{P(X)}{P(a)}$ is the same for all models, and the most probable model is the one that makes the data most probable.

When reliable information about the prior probability of a model is available we can use Bayes' law directly. This technique is known as *maximum a posteriori* model selection. For instance, over a wide variety of experiments, we may have observed that fair coins are far more common than unfair coins. It is very implausible that any particular coin would be unfair, but not impossible. While our prior knowledge should bias us toward the conclusion that a new coin is fair, it does not determine our conclusion. The likelihood of a model together with its prior probability can be used to determine which model has the highest probability of explaining the data. We want a model that both explains the data well and is plausible.

In general, evaluating joint probability over a large number of random variables is intractable. In practice most maximum likelihood schemes assume that the different trials of X are independent. The probability of co-occurrence is then the product of the independent RVs,

$$\ell(a) = \prod_{x_a \in a} P_X(x_a) \quad .$$

Maximizing $\ell(a)$ is still a daunting process. Significant simplification can be obtained by maximizing the logarithm of ℓ,

$$\log \ell(a) = \sum_{x_a \in a} \log P_X(x_a) \quad .$$

Log likelihood has the same maximum as $\ell(a)$, but has a much simpler derivative.

There is an interesting parallel between log likelihood and entropy. Recall that entropy is a statistic of X. The finite sample average of entropy, or empirical entropy, which will figure strongly later in the thesis, is

$$h_a(X) \equiv -E_a[\log P_X(X)] = -\frac{1}{N_a} \sum_{x_a \in a} \log P_X(x_a) \quad . \tag{2.25}$$

We can therefore conclude that

$$h_a(X) = -\frac{1}{N_a} \log(\ell(a)) \quad . \tag{2.26}$$

This provides us with an interpretation of model selection in terms of entropy. Instead of finding the model that makes the data most likely, we could instead find the model that has the lowest empirical entropy. Conversely, we could present a new interpretation of entropy: a distribution has low entropy if the probability of a sample drawn from that distribution is high. A distribution has high entropy if the sample has low probability. A density with a very narrow peak has low entropy because most of the samples will fall in the region where the density is large. A very broad density has high entropy because the samples are spread out and fall where the density is low.

The close relationship between entropy and log likelihood is well known, but is often overlooked by students of probability. As result the parallels between research on entropy and maximum likelihood can be easily missed. The fact that a system manipulates "entropy" does not make it necessarily any better, or different, than one based on likelihood. For instance, log likelihood model selection can be derived directly from the entropy framework using *cross entropy*. The cross entropy, $D(p_X \| p_{\tilde{X}})$, or asymmetric divergence is a measure of the difference between two distributions:

$$\begin{aligned}
D(p_X \parallel p_{\tilde{X}}) &= E_X\left[\log(\frac{p_X(X)}{p_{\tilde{X}}(X)})\right] && (2.27)\\
&= \int_{-\infty}^{\infty} \log\left(\frac{p_X(x')}{p_{\tilde{X}}(x')}\right) p_X(x')dx && (2.28)\\
&= \int_{-\infty}^{\infty} \log(p_X(x'))p_X(x')dx - \int_{-\infty}^{\infty} \log(p_{\tilde{X}}(x'))p_X(x')dx && (2.29)\\
&= -h(X) - E_X[\log(p_{\tilde{X}}(X))] && (2.30)\\
&\approx -h(X) - E_a[\log(p_{\tilde{X}}(X))] && (2.31)\\
&= -h(X) + h_a(X) \quad . && (2.32)
\end{aligned}$$

Cross entropy is non-negative, reaching zero if and only if p_X and $p_{\tilde{X}}$ are identical. Where maximum likelihood model selection searches for the model that makes the sample most likely, cross entropy model selection searches for the model that is closest, in the cross entropy sense, to the true distribution. If we use the approximation in (2.31), the two procedures are in fact identical. The first term in (2.31), $h(X)$, is a constant and does not play a role in model selection. The second term, $h_a(X)$, is $-N_a$ times the log likelihood of a sample drawn from X under the density $p(\tilde{x})$. Minimization of cross entropy implies the maximization of likelihood.

2.4 Modeling Densities

In this section we will describe a number of techniques for estimating densities from data. Understanding the process by which this is done is an important prerequisite for understanding the main algorithms in this thesis. We will begin with a discussion of the most widely observed continuous density: the Gaussian density[3]. We will then derive an closed form expression for the most likely Gaussian given a sample. This section will also include a discussion of other parametric density functions and finally a non-parametric technique for estimating densities known as Parzen window density estimation.

2.4.1 The Gaussian Density

The most ubiquitous of all random processes is the Gaussian or normal density. It literally appears everywhere. The most common justification arises from the "central limit theorem", which shows that the density of the sum of a very large number of independent random variables will tend toward Gaussian. An equally important justification is that the mathematics of the Gaussian density are quite simple because it is an exponential. Moreover, since any linear function of a Gaussian density is itself Gaussian, they are widely used in linear systems theory. It is almost certainly the case that the majority of continuous random processes have been modeled as Gaussians; whether they are or not. A *Gaussian density* is defined as:

$$g_{\psi}(x-\mu) \equiv \frac{1}{\sqrt{2\pi\psi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\psi}} \quad . \tag{2.33}$$

[3]For the sake of brevity we will sometimes refer to a Gaussian density as simply a Gaussian.

The parameters ψ and μ are the variance and mean of the density. One can demonstrate this with some clever integration.

The Gaussian density can also be defined in higher dimensions:

$$g_\psi(x-\mu) \equiv \frac{1}{(2\pi)^{\frac{n}{2}}|\psi|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu)^T\psi^{-1}(x-\mu)) \quad . \tag{2.34}$$

In a d dimensional space, the mean μ is a d-vector. The variance is replaced by a covariance matrix, a d-by-d matrix ($|\psi|$ is the determinant of ψ). Recall that variance is defined as the expected square of the difference from the mean; covariance is somewhat more complex:

$$\psi_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])] \quad , \tag{2.35}$$

where X_i is the i'th component of the RV X. The diagonal entries of ψ contain the variances of the components. The off-diagonal entries measure the expected co-variation.

Equation (2.33) defines an infinite family of density functions. Any one member of this family is determined by its mean and its variance. Before we can attempt to model an unknown density with a density from this family, we must first decide if the density is Gaussian. Maximum likelihood model selection can then be used to estimate μ and ψ from a sample. The form of the Gaussian density makes finding the maximum likelihood parameters very easy. The log likelihood of a sample of a Gaussian density is

$$\log(\ell(a)) = \sum_{x_a \in a} \log(P_X(x_a)) \tag{2.36}$$

$$= \sum_{x_a \in a} \log(g_\psi(x_a - \mu)) \tag{2.37}$$

$$= \sum_{x_a \in a} \log(\frac{1}{\sqrt{2\pi\psi}}) - \frac{1}{2}\frac{(x_a-\mu)^2}{\psi} \quad . \tag{2.38}$$

The most likely μ minimizes

$$\sum_{x_a \in a} (x_a - \mu)^2 \quad ,$$

a quadratic function of μ. Differentiating and solving for zeroes we find that the most likely μ is

$$\mu = \frac{1}{N_a} \sum_{x_a \in a} x_a \quad .$$

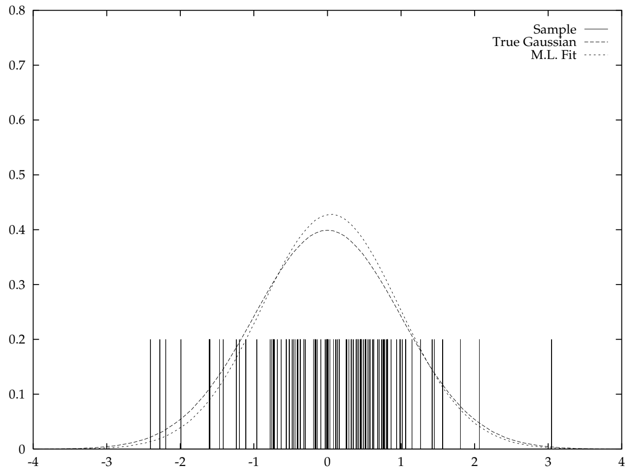


Figure 2.1: Three views of a Gaussian density with a mean of 0.0 and a variance 1.0: First a sample of 100 points drawn from the density. Each point is is represented a vertical black line. Second the density of the true Gaussian. Third the density of the Gaussian estimated from the sample (mean = 0.045, variance = 0.869).

This is a very satisfying result. We have proven that the most likely estimate for the mean, μ, is the mean of the sample. A very similar argument can be used to prove that the maximum likelihood estimate for the variance, ψ, is the sample variance:

$$\psi = \frac{1}{N_a} \sum_{x_a \in a} (x_a - \mu)^2 \quad .$$

Figure 2.1 displays a 100 point sample drawn from a Gaussian density. The true density is shown together with the most likely model. Because the sample mean and sample variance are not perfect measures of the true mean and variance, the most likely model is not perfect. The accuracy of the estimated mean and variance gets better as the sample size increases. Even for a sample of 100 points there is significant variability in the estimated model for different samples. Figure 2.2 shows ten different estimates from ten different samples of the same density.

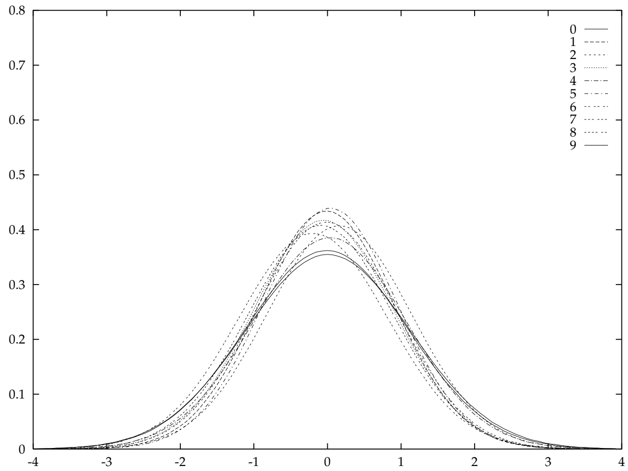


Figure 2.2: The maximum likelihood density estimates for ten different samples of 100 points drawn from the same Gaussian.

2.4.2 Other Parametric Densities

Finding the most likely Gaussian model for a sample is a very efficient operation. The mean and the variance are trivially computable in linear time. Efficient estimation is a property shared by all of the exponential densities, a class of densities which include the Gaussian density, the Gamma density, and others. For all other types of densities it is not possible to find maximum likelihood parameter estimates directly from statistics of the density. The most likely set of parameters must be determined by a search process. Since there are an infinite number of possible parameter values, finding values for these parameters that are optimal is not straightforward. Generally problems of this sort are solved using an iterative refinement process known as gradient descent. The gradient descent procedure is described in Appendix A.1.

The Gaussian density has many advantages. Why not use it to model every sample? The simple answer is that not all real densities are Gaussian. In fact, many real densities are far from Gaussian. One of the strongest limitations of the Gaussian, and of all the exponential densities, is that they are unimodal (they have a single peak). Modeling densities that have multiple peaks as if they had a single peak is foolhardy. Figure 2.3 shows an attempt to fit a two peaked function with a single Gaussian. In many situations it may seem as though the simplicity and efficiency that arises from using a Gaussian density outweigh the added accuracy that arises from using a more accurate model. As we will see, this is a temptation

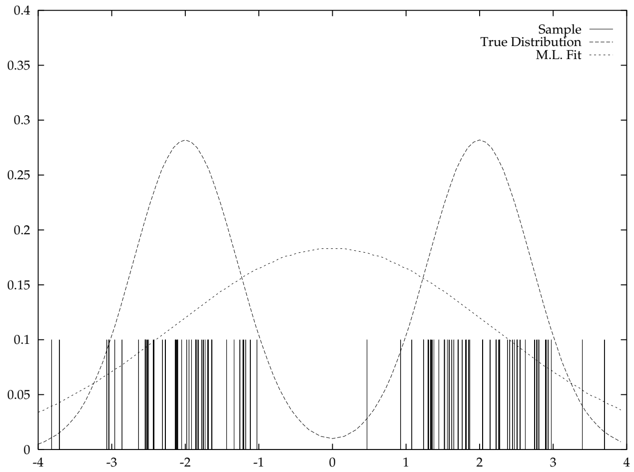


Figure 2.3: Three views of a density constructed from a combination of two Gaussians. The Gaussians have variance 0.3 and means of 2.0 and -2.0 respectively. As before the sample contains 100 points. The maximum likelihood Gaussian has mean 0.213 and variance 3.824.

to which many have succumbed.

Once the decision to use a more complex model has been made, the set of possible model densities is literally infinite. In terms of accuracy this is an unambiguous advantage. Density can be modeled by any function that can be guaranteed to integrate to one. The most common model after the simple Gaussian is a mixture of Gaussians:

$$M(x, W) = \sum_{i=1}^{N} c_i \; g_{\psi_i}(x - \mu_i) \quad , \tag{2.39}$$

here W represents the collection of parameters $(N, \{\mu_i\}, \{\psi_i\}, \{c_i\})$. When $\sum c_i = 1$, the mixture model is guaranteed to integrate to one. A mixture density need not be uni-modal; it may have as many as N peaks. Figure 2.3 contains a graph of a mixture of Gaussian density with two equal components. Given a large number of Gaussians, almost any density can be modeled accurately. As before, maximum likelihood can be used to select the best set of parameters for a given sample a. It is possible to search for the correct parameter vector using gradient ascent, but for Gaussian mixture models there is a more efficient technique known as Expectation Maximization (Dempster et al., 1977). In either case finding the best parameter vector can involve a lengthy search process.

While mixture models are fairly popular, almost any parameterized function can be used as a density estimate. Within the neural networks literature some have trained back propagation neural networks to approximate densities (Jacobs et al., 1991) (see also (Haykin, 1994) for an excellent review of neural network research). There is nothing terribly special about using a neural network for this purpose. It is just another form of parametric density estimation.

Now that we have some feel for the density estimation process, it is critical that important limitations be pointed out. Whenever one estimates a density from a sample a very important first step is required: assumptions must be made about the form of the density. The space of possible functions is so large that for any sample there are an infinite number of density functions that fit it equally well. Continuous density can defy intuition. For instance it is always possible to define a density that makes a given sample infinitely likely. Take for example a density function made up of delta functions. As we did in section 2.2 we could make up a function with a delta function for each trial:

$$p(X) = \frac{1}{N_a} \sum_{x_a \in a} \delta(x - x_a) \ . \tag{2.40}$$

The likelihood of this model density is then ∞, and is guaranteed to be bigger than any other density's likelihood. Wouldn't this imply that fitting any other density is sub-optimal? While intuition argues against such an artificial density, there is no principled scheme for dealing with this dilemma.

Much has been written about this problem in the machine learning literature, where it is called function approximation. There simply is not enough information in a finite sample to uniquely determine which of the infinite number of possible functions fit the sample best. The only solution is to make strong assumptions about the form of the correct function: for example that it is Gaussian, that it is polynomial, or that it is smooth. These assumptions provide a strong prior probability over the space of possible functions. Together the likelihood of a function and its prior probability can often uniquely determine a solution.

Maximum likelihood model selection is not guaranteed to do a good job of density estimation. There are three reasons why the most likely model may fail to be an accurate model. The first reason is that the set of evaluated models may not contain the correct model. This is called *inductive inadequacy* and it arises when the underlying assumptions about the density are wrong. The second reason is that maximum likelihood can be fooled.

An especially unlikely sample, as in our example of the unbiased coin (see Section 2.3), can lead to a model estimate that is not correct. This is a question of confidence. A larger sample is less likely to be unusual, and gives us more confidence in our model. The third reason is that the search through parameter space may fail. A good solution may exist but cannot be found, for example if there are local minima.

2.4.3 Parzen Window Density Estimation

The final class of density functions we will discuss are called non-parametric density estimators. For these models no search for parameters is needed. While parametric methods use the parameters as the model, non-parametric methods use the sample to directly define the model.

The non-parametric scheme on which we will focus is known as Parzen window density estimation. The general form of the density is:

$$P^*(x, a) \equiv \frac{1}{N_a} \sum_{x_a \in a} R(x - x_a) \ , \tag{2.41}$$

where a is a sample and R is a valid density function. The function R is often called the smoothing or window function. The quality of the approximation is dependent both on the functional form of R and its width. Different window functions will lead to very different density estimates. The Gaussian density is a common selection for R, making the Parzen density estimate a mixture of Gaussians. There is one Gaussian centered at each sample. Figure 2.4 contains a graph of a density, a sample, and the Parzen estimate constructed from the sample. Figure 2.5 contains a graph of ten different Parzen estimates from ten different samples. The different Parzen estimates do not show significantly more variation than the Gaussian estimates shown in Figure 2.2.

In practice the Parzen density estimate is much more flexible than a parametric density estimate. Where parametric techniques make very strong assumptions about the functional form of the density to be approximated, Parzen estimation requires only that the density be smooth. Figure 2.6 shows the Parzen density estimate of a bimodal distribution. Contrast it to the parametric estimate of the same density as shown in Figure 2.3.

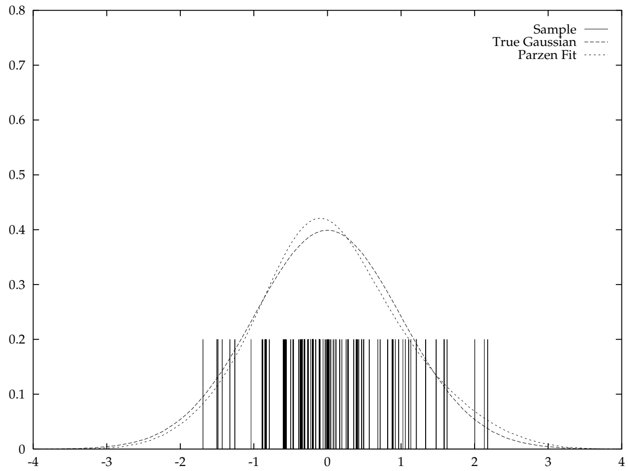


Figure 2.4: Three views of a Gaussian density with a mean of 0.0 and a variance 1.0: First a sample of 100 points drawn from the density. Each point is is represented a vertical black line. Second the density of the true Gaussian. Third the Parzen density estimate constructed from the sample. The window functions are Gaussians with variance 0.35.

Intuitively, the Parzen density estimator computes a local, or windowed, average of the sample. Looking back to (2.41), notice that if R is symmetrical about the origin we can view the window function as being centered on the query point, x, rather than at the data points. Viewed in this light, the density estimate at a query point is a weighted sum over the sample, where the weighting is determined by the window function. The most common window functions are unimodal, symmetric about the origin, and fall off quickly to zero. In effect, the window function defines a region centered on x in which sample points contribute to the density estimate. Points that fall outside of this window do not contribute. The density estimate at x is the ratio of the number of weighted sample points within the window divided by the total number of sample points, N_a. Getting a reliable estimate of this ratio involves having a reasonable number of points fall into the window around the query point. The number of points that we expect to fall into this window is a function both of the size of the sample and the size of the window. As the number of points that fall into a window decreases, the variance of the Parzen density estimate increases. We will analyze the variance of the Parzen estimate later in the chapter.

The balance of computation required by Parzen window density estimation is qualitatively very different from parametric schemes. Constructing a parametric model involves a lengthy search through parameter space that takes more time for larger samples. Constructing a

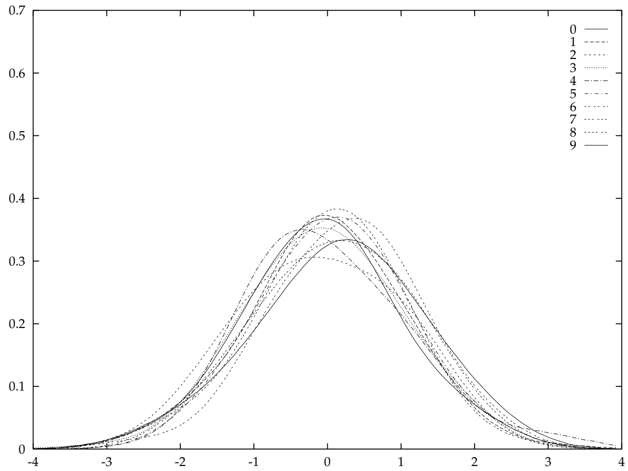


Figure 2.5: The Parzen density estimates for ten different samples of 100 points drawn from the same Gaussian density.

Parzen model is cheap. One need only memorize the sample. Evaluating a parametric model is usually efficient. Once the parameters are known the number of operations required is usually very small and does not grow with the size of the sample. Evaluating $P^*(x,a)$ is more expensive; requiring time proportional to the size of the sample. The overall computational complexity of either technique is a function of how they are used.

Though the Parzen estimate is a mixture model, it is not the maximum likelihood mixture model. Unlike the Parzen estimate, the maximum likelihood model is not constrained to place one Gaussians at each of the sample points. There is however an asymptotic proof of Parzen convergence that relies on the law of large numbers. The Parzen estimate can be written as a sample mean:

$$P^*(x_0,a) = \frac{1}{N_a}\sum_{x_a \in a} R(x_0 - x_a) = E_a[R(x_0 - X)] \ .$$

In the limit this equals the true expectation which in turn is a convolution

$$\lim_{N_a \to \infty} P^*(x,a) = E[R(x-X)] \tag{2.42}$$

$$= \int_{-\infty}^{\infty} R(x-x')p(x')dx' \tag{2.43}$$

$$= (R * p)(x) \ . \tag{2.44}$$

So $P^*(x,a)$ converges to $p(x)$ if and only if $p(x) = (R*p)(x)$. There are two distinct conditions under which equality holds. The first is that R tends toward the delta function

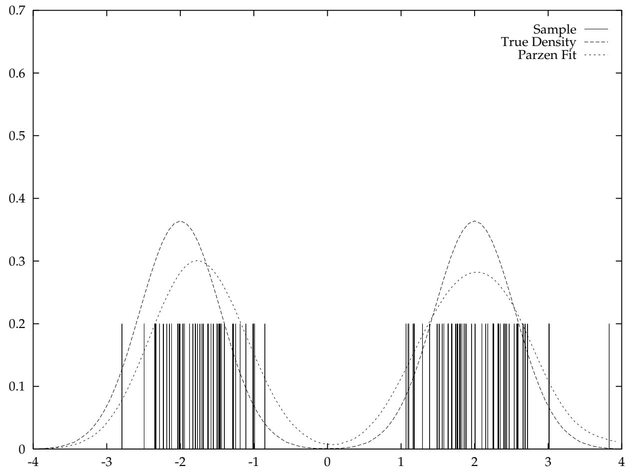


Figure 2.6: Three views of a density constructed from a combination of two Gaussians. The Gaussians have variance 0.3 and means of 2.0 and -2.0 respectively. As before the sample contains 100 points. The Parzen estimate is constructed with Gaussians of variance 0.20.

when the sample size approaches infinity. The second occurs when convolution by R does not change $p(x)$. In theory this could be achieved when $p(x)$ has bounded frequency content and R is a perfect low pass filter. In practice approximate equality holds whenever $p(x)$ has low frequency content and R is primarily a low pass filter, for example when $p(x)$ is a smooth function and R is a Gaussian. Finally, whenever $p(x) = (R * p)(x)$ the Parzen estimate, $P^*(x, a)$, is an unbiased estimator of $p(x)$.

There are other conditions under which the Parzen estimate will converge to the correct density estimate. This proof assumes that the samples are corrupted by measurement noise of a known density. Instead of X, a corrupted random variable, $\widetilde{X} = X + \eta$, is observed. If η were known the probability of X would be,

$$p(X = x|\widetilde{X} = \tilde{x}, \eta) = \delta(x - \tilde{x} - \eta) \ .$$

Without knowledge of η we must integrate over all its possible values,

$$\begin{aligned} p(X = x|\widetilde{X} = \tilde{x}) &= \int_{-\infty}^{\infty} p(X = x|\widetilde{X} = \tilde{x}, \eta')p_\eta(\eta')d\eta' && (2.45) \\ &= \int_{-\infty}^{\infty} \delta(x - \tilde{x} - \eta)p_\eta(\eta)d\eta' && (2.46) \\ &= p_\eta(x - \tilde{x}) \ . && (2.47) \end{aligned}$$

To compute $p(X)$ we must integrate over $\widetilde{X}$. We show that the integral is approximated by the Parzen estimator,

$$
\begin{aligned}
p(X = x) &= \int_{-\infty}^{\infty} p(X = x | \widetilde{X} = \widetilde{X}') p_{\widetilde{X}}(\widetilde{X}') d\widetilde{X}' && (2.48)\\
&= \int_{-\infty}^{\infty} p_\eta(x - \widetilde{X}') p_{\widetilde{X}}(\widetilde{X}') d\widetilde{X}' && (2.49)\\
&= E_{\widetilde{X}}[p_\eta(x - \widetilde{X})] && (2.50)\\
&\approx E_a[p_\eta(x - \tilde{x}_a)] && (2.51)\\
&= \frac{1}{N_a} \sum_{\tilde{x}_a \in a} p_\eta(x_0 - \tilde{x}_a) \ , && (2.52)
\end{aligned}
$$

where a is a sample of $\widetilde{X}$. The probability of the uncorrupted random variable X is approximated by the Parzen estimate constructed from the samples of $\widetilde{X}$ where the smoothing function is the density function of the noise. The probability of the corrupted random variable can be derived from a very similar argument,

$$
p(\widetilde{X} = \tilde{x}) \approx \frac{1}{N_a} \sum_{\tilde{x}_a \in a} (p_\eta * p_\eta)(x_0 - \tilde{x}_a) \ .
$$

The probability of a noise corrupted random variable $\widetilde{X}$ is approximated by the Parzen estimate using the smoothing function $(p_\eta * p_\eta)(x)$. This result is independent of the density of X. Often η is Gaussian noise, a very common assumption that we will return to in our discussions of entropy. The smoothing function is then a Gaussian density that has twice the standard deviation of η.

Finding the Best Smoothing Functions

As we have seen, when a priori information about the density is available Parzen estimation will converge to the correct density. Moreover, when we know either that the density is smooth or that it has been perturbed by noise it is possible to find the correct smoothing function. In the absence of a priori information, the quality of the Parzen estimate is dependent on the variance ψ of the smoothing functions. Figures 2.7 and 2.8 display the dependence of the density estimate on ψ. Each shows the Parzen estimates computed from a 100 point sample as ψ is changed. Notice that the actual density function that results is very dependent on the variance. The qualitative nature of this dependence varies across the range of variances

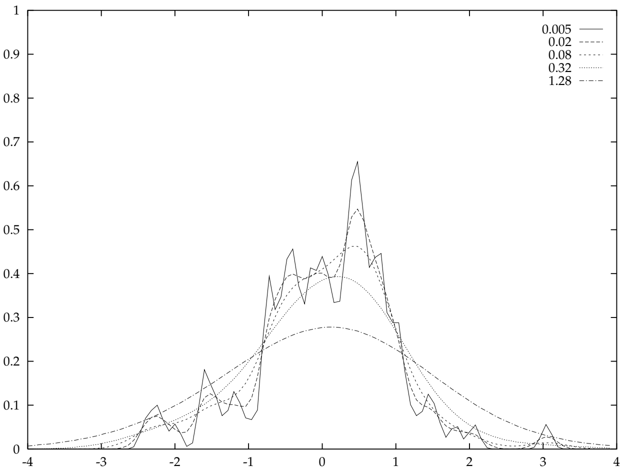


Figure 2.7: Five plots of the Parzen density estimates derived from a 100 point sample of a Gaussian. The Gaussian has variance 1.0 and mean 0.0. The different estimates use a different value for the variance of the component smoothing functions. The variances used range over a factor of 256, from 0.005 to 1.28.

shown. When the variance of the smoothing function is small, less that 0.1, the resulting density changes very rapidly as variance is changed. Above 0.2 small changes in variance do not change the resulting density nearly as rapidly.

Selection of the correct variance for the smoothing functions need not be a hit or miss process. Much in the same way that likelihood can be used to find the parameters of a Gaussian to fit a sample, likelihood can be used to find the variance of the Gaussians that make up the Parzen estimate. In general it is possible to compute the best variance for each Gaussian in the Parzen density estimate separately. This process requires a great deal of time and data. Since we wish to preserve the simplicity of the Parzen estimate, a single variance will be used for all of the smoothing functions.

Recall that likelihood is maximized when empirical entropy is minimized (see Section 2.3.1). Since subsequent chapters will focus on empirical entropy, we will use empirical entropy to estimate the optimal variance. Figure 2.9 graphs the empirical entropy of the sample versus variance. The sample used in this graph is the same as was used to estimate the densities in Figures 2.7 and 2.8. The broad minimum in entropy at 0.25 implies that the Parzen density estimate is not critically dependent on variance. The variance need only be within a factor of ten of the optimal variance.

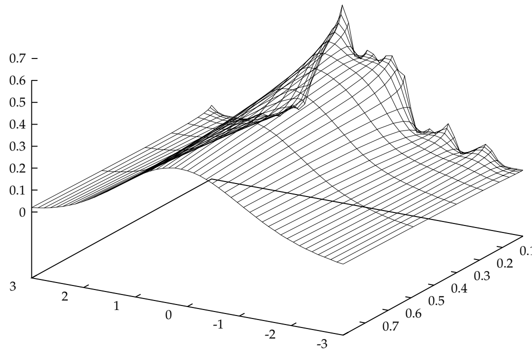


Figure 2.8: A parametric surface plot of the Parzen density versus variance (this is the same data shown in the previous graph). The horizontal and vertical axes are the location and density respectively. Variance changes with depth in the graph. Here variance ranges from 0.80 to 0.01.

The true entropy of a Gaussian with variance 1.0 is 1.419. The optimal Parzen density estimate has an empirical entropy of 1.47. This close agreement is not coincidence. It is argued in the next chapter that the true entropy of a density can be effectively estimated from a Parzen density estimate.

There is a small technical note that should not be overlooked. We must be careful whenever the same sample that is used both to construct the Parzen estimate and to estimate entropy. Recall that the most likely, or lowest entropy, density estimate for a sample is a collection of delta functions centered at each point from the sample (see 2.2.1). We also know that this delta function density will have an entropy of negative infinity. The Parzen density is very similar in form to the delta function density. It too centers a function at each point from the sample. In the limit as the variance of the smoothing functions tends towards zero, each smoothing function approximates a delta function. Therefore the minimum empirical entropy should be obtained when the variance is zero. This difficulty only arises when the sample from which the density is estimated is the same as the sample with which empirical entropy is calculated. If these two samples are different, and the density is not degenerate in some way, then no point should appear in both samples. As the variance of the smoothing functions tends to zero the density at all points that are not in the Parzen sample tends to zero. As a result the empirical entropy should tend toward positive infinity as the variance

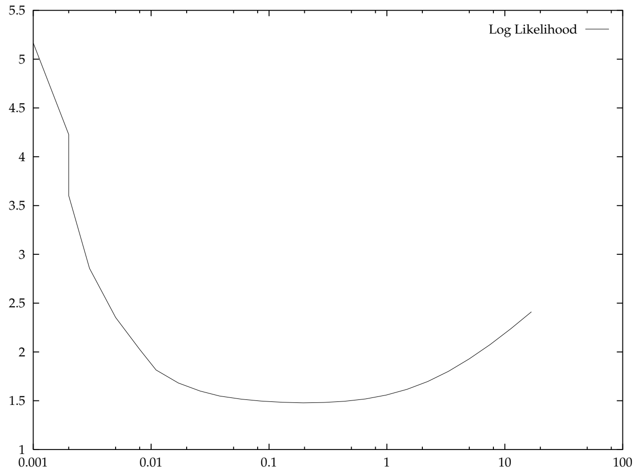


Figure 2.9: A log plot of negative log likelihood versus ψ. Near the minimum, the log likelihood is not terribly sensitive to the ψ. Values within a factor of 10 are all roughly equivalent.

tends to zero. This effectively precludes the solution where the variance of the smoothing functions is zero.

We can simulate having two different samples by a process called *cross-validation*. Cross-validation splits a single sample a into two samples. One sample has a single point $\{x\}$ and the second contains the remaining points $a - \{x\}$. There are N_a different ways to split the sample in two parts. Rather than draw two different samples, we use the N_a different split samples. In each case the larger sample, of size $N_a - 1$, is used for the Parzen estimate, and the smaller sample is used to estimate the entropy. Estimating log likelihood or empirical entropy with two samples a and b yields

$$\log(\ell(b)) = -N_b h_b(X) = -N_b E_b[\log P^*(X, a)] \ , \tag{2.53}$$

versus

$$-N_a E_a[\log(P^*(x_a, a - \{x_a\}))] \ , \tag{2.54}$$

using cross validation. The cross validated empirical entropy is an unbiased estimate of the two sample empirical entropy.

The Quality of the Parzen Estimate

One way to evaluate the quality of the Parzen estimate is to evaluate the standard deviation of its estimate. Another, perhaps more useful statistic is the standard deviation normalized by the mean

$$\frac{\sigma(X)}{E[X]} \ . \tag{2.55}$$

The normalized standard deviation measures expected deviation from the mean as a function of the overall scale. For many types of problems, when the mean of a variable is large, small deviations about the mean are usually unimportant. But when the mean is very small, a small deviation can make a big difference. Normalized standard deviation is a good measure to use when the log of a variable is important (like log likelihood and entropy). Using the constant and linear terms of the Taylor expansion of logarithm and assuming that the standard deviation of X is small,

$$\sigma(\log(X)) \approx \frac{\sigma(X)}{E[X]} \ . \tag{2.56}$$

The standard deviation of the Parzen density estimate at a point x is a function of the total number of sample points used to estimate the density. The normalized standard deviation of a Parzen estimate is

$$\frac{\sigma(P^*(x,a))}{E[P^*(x,a)]} = \frac{\sigma(P^*(x,a))}{p(X)} \ , \tag{2.57}$$

where both the standard deviation and the expectation are taken over the space of possible samples. The two equations are equal whenever $P^*(x,a)$ is an unbiased estimator for $p(X)$.

The standard deviation of the Parzen estimate can be computed exactly when the smoothing functions are box car functions. The Parzen estimate is then the number of sample points that fall into the box car window divided by the total number of sample points,

$$P^*(x,a) \equiv \frac{1}{N_a} \sum_{x_a \in a} R(x - x_a) = \frac{kN_{in}}{N_a} \ . \tag{2.58}$$

where N_{in} is the number of points for which $R(x - x_a)$ is non-zero, and $k = R(0)$ is chosen so that P^* integrates to one. The standard deviation of the Parzen estimate is,

$$\sigma(P^*(x,a)) = \frac{k}{N_a}\sigma(N_{in}) \ . \tag{2.59}$$

Assuming each point in the sample is independent then

$$\sigma(P^*(x,a)) = \frac{k}{N_a}\sqrt{N_a P^*(x,a)\left(1 - \frac{P^*(x,a)}{k}\right)} \ . \tag{2.60}$$

The normalized standard deviation of the Parzen estimate is then

$$\frac{\sigma(P^*(x,a))}{E[P^*(x,a)]} = \frac{1}{E[P^*(x,a)]}\frac{k}{N_a}\sqrt{N_a P^*(x,a)\left(1 - \frac{P^*(x,a)}{k}\right)} \tag{2.61}$$

$$\approx \frac{1}{P^*(x,a)}\frac{k}{N_a}\sqrt{N_a P^*(x,a)(1 - \frac{P^*(x,a)}{k})} \tag{2.62}$$

$$\approx \frac{k}{\sqrt{N_a}}\sqrt{\frac{(1 - \frac{P^*(x,a)}{k})}{P^*(x,a)}} \ . \tag{2.63}$$

The Parzen estimate has a larger normalized standard deviation at points where the estimated probability is small. The Parzen density estimate converges to the true estimate at a rate proportional to $\frac{1}{\sqrt{N_a}}$.

The definition of Parzen window estimation can be generalized to higher dimensions by replacing the one dimensional smoothing functions by their d dimensional counterparts (see Section 2.4 for a d dimensional Gaussian). Though the definition of Parzen estimation is the same for any number of dimensions, the behavior of the algorithm can be very different. As the number of dimensions grows the number of data points required rapidly increases. In d dimensions, the window of support of a Gaussian smoothing function is an d dimensional sphere whose radius r is a function of its standard deviation. The volume of a d dimensional sphere of radius r is $V_d r^d$, where V_d is a constant dependent only on the dimension. Assuming that all of the sample data is contained in a sphere of radius R, on average $N_a\left(\frac{r}{R}\right)^d$ data points will fall in a randomly chosen window. Generally, r is selected so that $\frac{r}{R} < 1$. As a result with increased dimension the number of points falling in a randomly chosen window drops exponentially and the normalized standard deviation of P^* will increase rapidly. This implies that the Parzen density estimate will become very unreliable as dimensionality increases.

While in theory this could be remedied by increasing the size of the sample exponentially, things rapidly get out of hand. Empirical evidence argues against using Parzen estimation in many more than six dimensions.

Chapter 3

Empirical Entropy Manipulation and Stochastic Gradient Descent

This chapter presents a novel technique for evaluating and manipulating the empirical entropy of a distribution, called EMMA. The theory of entropy manipulation plays a critical role in the rest of this thesis and forms the algorithmic core in all of the applications.

There are a number of existing techniques that manipulate the entropy of a density. They each have significant theoretical and practical limitations that make them unsuitable for our purposes. We will begin with a description of these techniques, and two simple applications. The second part of the chapter describes a new procedure for evaluating empirical entropy, EMMA. We will present an efficient stochastic gradient scheme for extremizing the EMMA estimates. This scheme has applications outside of entropy manipulation.

The final section of this chapter presents a tutorial application of EMMA. We will show how EMMA can be used to derive an information theoretic version of principal components analysis.

3.1 Empirical Entropy

As we saw in the previous chapter, in many cases the true density of a random variable is not known. Instead one must make do with an estimate of the density obtained from a sample of the variable. Likewise, there is no direct procedure for evaluating entropy from a sample. A common approach is to first model the density from the sample, and then estimate the entropy from the density. This divides the problem into manageable parts which can be solved separately.

By far the most popular density model for entropy calculations is the Gaussian. This is based on two considerations: (1) finding the Gaussian that fits the data best is very easy (see Section 2.3.1) and (2) the entropy of the Gaussian can be directly calculated from its variance. The entropy of a Gaussian distribution is

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{\infty} g_\psi(x-\mu)\log(g_\psi(x-\mu))dx && (3.1)\\
&= -\int_{-\infty}^{\infty} g_\psi(x-\mu)\left[\log\left(\frac{1}{\sqrt{2\pi\psi}}\right) - \frac{1}{2}\frac{(x-\mu)^2}{\psi}\right]dx && (3.2)\\
&= E\left[\frac{1}{2}\frac{(x-\mu)^2}{\psi}\right] + \frac{1}{2}\log(2\pi\psi) && (3.3)\\
&= \frac{1}{2} + \frac{1}{2}\log(2\pi\psi) && (3.4)\\
&= \frac{1}{2}\log(e) + \frac{1}{2}\log(2\pi\psi) && (3.5)\\
&= \frac{1}{2}\log(2e\pi\psi) \ . && (3.6)
\end{aligned}$$

The entropy of a Gaussian is a function of its variance and is not a function of its mean. Wider Gaussians have more entropy and narrower Gaussians have less. There is a simple procedure for finding the empirical entropy of a Gaussian distribution: compute the variance of the sample and evaluate (3.6).

The equivalence between the log of variance and entropy can be used to reformulate well known signal and image processing problems as entropy problems. Since the logarithm is a monotonically increasing function, any technique that maximizes or minimizes the variance of a signal can be viewed as an entropy technique. Examples include principal components analysis, where variance is maximized, and least square solutions to matching problems, where variance is minimized. There is however one significant caveat. Variance maximization is only equivalent to entropy maximization if the density of the signal involved is Gaussian. When this assumption is violated it is possible to reduce entropy as variance is increased. All but a very few of the techniques that manipulate the entropy of a signal assume that the signals are Gaussian or another exponential distribution (Linsker, 1988; Becker and Hinton, 1992; Bell and Sejnowski, 1995). We will discuss these techniques in Section 7.

Principal Components Analysis

There are a number of signal processing and learning problems that can be formulated as entropy maximization problems. One well known example is *principal components analysis.* The principal component of a d dimensional distribution is a d dimensional vector. Given a density X every vector v defines a new random variable, $Y_v = X \cdot v$. The variance along an axis v is the variance of this new variable:

$$Var(Y_v) = E_X[(X \cdot v - E_X(X \cdot v))^2] \ . \tag{3.7}$$

The principal component v is the vector for which $Var(Y_v)$ is maximized.

In practice neither the density of X nor Y_v is known. The projection variance is computed from a sample a of points from X,

$$Var(Y_v) \approx Var_a(Y_v) \equiv E_a[(X \cdot v - E_a[X \cdot v])^2] \ . \tag{3.8}$$

We can then define a vector cost function,

$$C(v) = -Var_a(Y_v) \ , \tag{3.9}$$

which is minimized by the principal component vector. There are many schemes for finding the principal component of a density. One of the more elegant accomplishments of linear algebra is the proof that the first eigenvector of the covariance matrix of X, ψ_X, is the principal component vector.

Under the assumption that X is Gaussian we can prove that the principal component is the projection with maximum entropy. First, every projection of a Gaussian is also Gaussian.

Second, the entropy of a Gaussian is monotonically related to variance. Therefore, the Y_v corresponding the axis of highest variance is a Gaussian with the highest possible entropy. Moreover, principal components analysis finds the axis that contains the most information about X. The mutual information between X and Y_v is

$$I(X, Y_v) = h(Y_v) - h(Y_v|X) \quad . \tag{3.10}$$

This equation has two components. The first implies that Y_v will give you more information about X when Y_v has a lot of entropy. The second can be misleading. Since knowing X removes all of the randomness from Y_v, $h(Y_v|X)$ is negative infinity. This is not particularly bothersome precisely because relative entropy is relative. Only the differences between relative entropies are significant. The variable Y_{v_1} yields more information about X than Y_{v_2} when

$$I(X, Y_{v_1}) > I(X, Y_{v_2}) \tag{3.11}$$

$$h(Y_{v_1}) - h(Y_{v_1}|X) > h(Y_{v_2}) - h(Y_{v_2}|X) \tag{3.12}$$

$$h(Y_{v_1}) > h(Y_{v_2}) \quad . \tag{3.13}$$

We can conclude that the principal component axis carries more information about a distribution than any other axis.

Function Learning

There are other well known problems that can be formulated within the entropy framework. Let us analyze a simple learning problem. Given a random variable X we can define a functionally dependent RV, $Y = F(X \cdot v) + \eta$, which we assume has been perturbed by measurement noise. We are given samples of the joint occurrences of the RVs: $a = [...\{x_a, y_a\}...]$. How can we estimate v? Typically this is formulated as a least squares problem. A cost is defined as the sum of the squares of the differences between predicted $\hat{y}_a = F(x_a \cdot \hat{v})$ and the actual samples of Y,

$$C(\hat{v}) = \sum_{\{x_a, y_a\} \in a} (y_a - F(x_a \cdot \hat{v}))^2 \quad . \tag{3.14}$$

The cost is a function of the estimated parameter vector $\hat{v}$. The sum of squared difference can be justified from a log likelihood perspective (see Section 2.3.1). If we assume that the noise added to Y is Gaussian and independent of Y then the log likelihood of $\hat{v}$ is:

$$
\begin{aligned}
log(\ell(\hat{v})) &= \sum_{\{x_a, y_a\} \in a} log(p(Y = y_a | \hat{Y} = F(x_a \cdot \hat{v}))) && (3.15) \\
&= \sum_{\{x_a, y_a\} \in a} log(g_\psi(y_a - F(x_a \cdot \hat{v}))) && (3.16) \\
&= - \sum_{\{x_a, y_a\} \in a} (y_a - F(x_a \cdot \hat{v}))^2 + k \ , && (3.17)
\end{aligned}
$$

where k is a constant value independent of $\hat{v}$. The $\hat{v}$ that minimizes the sum of the squares is also the $\hat{v}$ that makes the data observed most likely. For most problems like this, gradient descent is used to search for $\hat{v}$.

Finally we can show that minimizing cost maximizes mutual information. We showed in Section 2.3.1 that log likelihood is related to sample entropy:

$$
log(\ell(\hat{v})) = -\frac{1}{N_a} h_a(Y|\hat{Y}) \approx -\frac{1}{N_a} h(Y|\hat{Y}) \ . \tag{3.18}
$$

The mutual information between Y and $\hat{Y}$ is

$$
I(Y, \hat{Y}) = h(Y) - h(Y|\hat{Y}) \ . \tag{3.19}
$$

The first term is not a function of $\hat{v}$. The second is an approximation of log likelihood—minimizing $C(\hat{v})$ maximizes $I(Y, \hat{Y})$.

Non-Gaussian Densities

There is a commonly held misconception that mutual information techniques are all equivalent to simple well known algorithms. Contrary to the impression that the above two examples may give, this is far from the truth. Entropy is *only* equivalent to least squares when the data is assumed to be Gaussian. The approach to alignment and bias correction that we will describe in the next chapters does not and could not assume that distributions were Gaussian. We will show that if the data were Gaussian our alignment technique would reduce to correlation.

There are a number of non-Gaussian problems that can be solved using entropy or mutual information. Bell has shown that signal separation and de-correlation can be thought of as entropy problems (Bell and Sejnowski, 1995). Bell's technique can be derived both for Gaussian and non-Gaussian distributions. Bell shows that the Gaussian assumption leads to a well known and ineffective algorithm. When the signals are presumed to be non-Gaussian, the resulting algorithms are much more effective. Many compression and image processing problems clearly involve non-Gaussian distributions.

In theory, empirical entropy estimation can be used with any type of density model. The procedure is the same: estimate the density from a sample and compute the entropy from the density. In practice, the process can be computationally intensive. The first part, maximum likelihood density estimation, is an iterative search through parameter space. The second, evaluating the entropy integral, may well be impossible. For example, there is no known closed form solution for the entropy of a mixture of Gaussians. The entropy integral can however be approximated as a sample mean:

$$h(X) \approx h_b(X) = E_b[log(\hat{p}(X))] \ , \tag{3.20}$$

where $E_b[\ ]$ is the sample mean taken over the sample b, $\hat{p}()$ is the estimate for the sample density and $h_b(X)$ is the sample entropy first introduced in Section 2.3.1. The sample mean converges toward the true mean at a rate proportional to $1/\sqrt{N_b}$ (N_b is the size of b). Based on this insight, two samples can be used to estimate the entropy of a distribution: the first is used to estimate the density, the second is used to estimate the entropy.

While the two sample approach can be used to estimate entropy, it is not a practical algorithm for entropy manipulation. In the two applications above, changes in the parameter vector effect the densities that are being approximated. As the search through parameter space adjusts the parameter vector a new sample must be drawn, a new density estimated, and the derivative of entropy evaluated. If estimating the density is itself a complex search process, the search for the correct parameter vector can take an unbearably long time.

3.2 Estimating Entropy with Parzen Densities

In this section we will describe a technique that can effectively estimate and manipulate the entropy of non-Gaussian distributions. The basic insight is that rather than use maximum likelihood to estimate the density of a sample, we will instead use Parzen window density estimation (see Section 2.4.3). The Parzen scheme for estimating densities has two significant advantages over maximum likelihood: (1) since the Parzen estimate is computed directly from the sample, there is no search for parameters; (2) the derivative of the entropy of the Parzen estimate is simple to compute.

In the following general derivation we will assume that we have samples of a random variable X, and we would like to manipulate the entropy of the random variable $Y = F(X, v)$. The entropy, $h(Y)$, is now a function of v and can be manipulated by changing v. Since there is no direct technique for finding the parameters that will extremize $h(Y)$ we will search the parameter space using gradient descent. The following derivation assumes that Y is a vector random variable. The joint entropy of a two random variables, $h(Y_1, Y_2)$, can be evaluated by constructing the vector random variable, $W = [Y_1, Y_2]^T$ and evaluated $h(W)$.

The form of the Parzen estimate constructed from a sample a is

$$P^*(y, a) = \frac{1}{N_a} \sum_{y_a \in a} g_\psi(y - y_a) \ , \tag{3.21}$$

where the Parzen estimator is constructed with Gaussian smoothing functions. Given $P^*(y, a)$ we can approximate entropy as the sample mean,

$$\begin{aligned} h^*(Y) &\equiv E_b[log(P^*(Y, a))] && (3.22) \\ &= \frac{1}{N_b} \sum_{y_b \in b} log(P^*(y_b, a)) \ , && (3.23) \end{aligned}$$

computed over a second sample b ($h^*(X)$ is the EMMA estimate of empirical entropy).

In order to extremize entropy we must calculate the derivative of entropy with respect to v. This may be expressed as

$$\frac{d}{dv} h^*(Y) = \frac{-1}{N_b} \sum_{y_b \in b} \frac{\sum_{y_a \in a} \frac{d}{dv} g_\psi(y_b - y_a)}{\sum_{y_a \in a} g_\psi(y_b - y_a)} \ , \tag{3.24}$$

and, after differentiating the Gaussian,

$$\frac{d}{dv} h^*(Y)) = \frac{1}{N_b} \sum_{y_b \in b} \frac{\sum_{y_a \in a} g_\psi(y_b - y_a)\ (y_b - y_a)^T \psi^{-1} \frac{d}{dv}(y_b - y_a)}{\sum_{y_a \in a} g_\psi(y_b - y_a)} \ . \tag{3.25}$$

This expression may be written more compactly as follows,

$$\frac{d}{dv}h^*(Y) = \frac{1}{N_b}\sum_{y_b\in b}\sum_{y_a\in a} W_y(y_b, y_a)\ (y_b - y_a)^T\psi^{-1}\frac{d}{dv}(y_b - y_a)\ \ , \tag{3.26}$$

using the following definition:

$$W_y(y_b, y_a) \equiv \frac{g_\psi(y_b - y_a)}{\sum_{y_a\in a} g_\psi(y_b - y_a)}\ \ . \tag{3.27}$$

$W_y(y_b, y_a)$ takes on values between zero and one. It will approach one if y_b is significantly closer to y_a than any other element of a. It will be near zero if some other element of a is significantly closer to y_b. Distance is interpreted with respect to the squared Mahalonobis distance (see (Duda and Hart, 1973))

$$D_\psi(y) \equiv y^T\psi^{-1}y\ \ .$$

Thus, $W_y(y_b, y_a)$ is an indicator of the degree of match between its arguments, in a "soft" sense. It is equivalent to using the "softmax" function of neural networks (Bridle, 1989) on the negative of the Mahalonobis distance to indicate correspondence between y_b and elements of a.

Equation 3.26 may also be expressed as

$$\frac{d}{dv}h^*(Y) = \frac{1}{N_B}\sum_{y_b\in b}\sum_{y_a\in a} W_y(y_b, y_a)\ \frac{d}{dv}\frac{1}{2}D_\psi(y_b - y_a)\ \ . \tag{3.28}$$

In this form it is apparent that to reduce entropy, the parameters v should be adjusted such that there is a reduction in the average squared distance between points which W indicates are nearby.

Before moving on, it is worth reemphasizing that for most density models $\frac{d}{dv}h(Y)$ is very difficult to compute. The general derivation of the derivative of entropy is much more complex than the Parzen derivation:

$$\frac{\partial h(Y)}{\partial v} \approx \frac{d}{dv}\frac{1}{N_b}\sum_{y_b\in b} logp(y_b, a) \tag{3.29}$$

$$= \frac{1}{N_b} \sum_{y_b \in b} \frac{\frac{d}{dv} p(y_b, a)}{p(y_b, a)} \tag{3.30}$$

$$= \frac{1}{N_b} \sum_{y_b \in b} \frac{\frac{d}{dy_b} p(y_b, a) \cdot \frac{dy_b}{dv} + \frac{d}{da} p(y_b, a) \cdot \frac{da}{dv}}{p(y_b, a)} \quad . \tag{3.31}$$

The numerator of the derivative has two components. The first, $\frac{d}{dy_b} p(y_b, a) \cdot \frac{dy_b}{dv}$, is the change in entropy that results from changes in the sample b. The second, $\frac{d}{da} p(y_b, a) \cdot \frac{da}{dv}$, is far more problematic. The second component is a measure of the change in the density estimate that results from changes in the sample a. In the Parzen framework the two components of the derivative collapse into a single term and can be directly computed from the samples. In the maximum likelihood framework $p(y_b, a)$ is a complex function of the sample. Since there is no closed form function that computes the density estimate from the sample, computing its derivative can be very difficult.

3.3 Stochastic Maximization Algorithm

The variance maximization/minimization applications described above (principal components analysis and learning) are deterministic procedures. Starting from an initial guess, gradient descent uses the derivative of cost to repeatedly update the parameter vector. Two different runs that start from the same initial parameters will end up with the same final parameters. Our justification for using probability and entropy to analyze these problems is purely convenience. There is nothing random about these problems once the samples are drawn. One of the benefits of understanding the probabilistic interpretation of these problems is that we can introduce randomness into our solutions and understand its effect. Here is a simple example: we want to know the average of a large sample of data. Without knowing anything else, it would make sense to sum over the entire sample. But, if we needed only a rough estimate of the average, significant computation could be saved by averaging over a subset of the sample. Furthermore, knowledge of the sample variance would allow us to compute the size of the subsample needed to estimate the mean to a given precision.

A similar analysis can be applied to principal components analysis or function learning. The cost of a particular parameter vector is computed by summing over an entire sample as we did in Equation 3.14. But, when that sample is very large this expectation can be approximated by a smaller random sample. The same argument applies to the gradient. Since the gradient is defined as an average over a very large sample it may make sense to approximate it over a smaller random sample. When we use random samples, both the error estimate and the gradient estimate are now truly random. For very large samples, accurate error/gradient estimates can be made without averaging over the entire sample. For problems where the gradient needs to be evaluated often, this can save significant computation.

Though a random estimate of the gradient is cheaper to compute, it could be useless. Under what conditions does it make sense to use a random gradient estimate? The theory of stochastic approximation tells us that stochastic estimates of the gradient can be used instead of the true gradient when the following conditions hold: (1) the gradient estimate is unbiased; (2) the parameter update rate asymptotically converges to zero; (3) the error surface is quadratic in the parameters (Robbins and Munroe, 1951; Ljung and Söderström, 1983; Haykin, 1994). The first condition requires that on average the estimate for the gradient is the true gradient. The second insures that the search will eventually stop moving about randomly in parameter space. In practice the third condition can be relaxed to include most smooth non-linear error surfaces; though there is no guarantee that the parameters will end up in any particular minimum.

Returning our attention to equations (3.22) and (3.26), notice that the both the calculation of the EMMA entropy estimate and its derivative involve a double summation. One summation is over the points in sample a and another over the points in b. As a result the cost of evaluation is quadratic in sample size: $O(N_a N_b)$. We will present an experiment where the derivative of entropy for an image containing $60,000$ pixels is evaluated. While the true derivative of empirical entropy could be obtained by exhaustively sampling the data, a random estimate of the entropy can be obtained with much less computation. This is especially critical in entropy manipulation problems, where the derivative of entropy is evaluated many thousands of times. Without the quadratic savings that arise from using smaller samples entropy manipulation would be impossible.

For entropy manipulation problems involving large samples we will use stochastic gradient descent. Stochastic gradient descent seeks a local maximum of entropy by using a stochastic estimate of the gradient instead of the true gradient. Steps are repeatedly taken that are proportional to the approximation of the derivative of the mutual information with respect to the parameters:

Repeat:

$$
\begin{aligned}
a &\leftarrow \{N_a \text{ samples drawn from } y\} \\
b &\leftarrow \{N_b \text{ samples drawn from } y\} \\
v &\leftarrow v + \lambda \frac{dh}{dv}
\end{aligned}
$$

where $\frac{dh}{dv}$ is the derivative of entropy evaluated over samples a and b, v is the parameter to be estimated, and the parameter λ is called the "learning rate". The above procedure is repeated a fixed number of times or until convergence is detected. In most problems the initial value of λ is reduced during the search. In subsequent chapters we will describe experiments where samples of 50 or less can be used to effectively find entropy maxima.

Stochastic approximation does not seem to be well known in the computer vision community. We believe that it is applicable to a number of cost minimization problems that arise in computer vision. Stochastic gradient descent is most appropriate for tasks where evaluation of the true gradient is expensive, but an unbiased estimate of the gradient is easy to compute. Examples include cost functions whose derivative is a sum over all of the pixels in an image. In these cases, stochastic gradient search can be orders of magnitude faster than even the most complex second order gradient search schemes. In the experimental section (Chapter 6) of this thesis we will briefly describe joint work where an existing vision application was sped up by a factor of fifty using stochastic approximation.

Convergence of Stochastic EMMA

Most of the conditions that insure convergence of stochastic gradient descent are easy to obtain in practice. For example, it is not really necessary for the learning rate to asymptotically converge to zero. At non-zero learning rates the parameter vector will move randomly about the minimum/maximum endlessly. Smaller learning rates make for smaller excursions from the true answer. An effective way to terminate the search is to detect when on average the parameter is not changing and then reduce the learning rate. The learning rate only needs to approach zero if your goal is zero error, something that no practical system can achieve anyway. A better idea is to reduce the learning rate until the parameters have a reasonable variance and then take the average parameters.

The first proofs of stochastic approximation required that the error be quadratic in the parameters. More modern proofs are more general. For convergence to a particular optimum, the parameter vector must be guaranteed to enter that optimum's basin of attraction infinitely often. A basin of attraction of an optimum is defined with respect to true gradient descent. Each basin is a set of points from which true gradient descent will converge to the same optimum. Quadratic error surfaces have a single optimum and the basin of attraction is the entire parameter space. Non-linear error spaces may have many optima, and the parameter space is partitioned into many basins of attraction. When there is a finite number of optima, we can prove that stochastic gradient descent will converge to one of them. The proof proceeds by contradiction. Assume that the parameter vector never converges. Instead it wanders about parameter space forever. Since parameter space in partitioned into basins of attraction it is always in some basin. But since there are a finite number of basins it must be in one basin infinitely often. So it must converge to the optimum of that basin.

One condition will give us more trouble than the others. The stochastic estimate of the gradient must be unbiased. It is *not* true that the sample approximation for empirical entropy is unbiased. Moreover, we have been able to prove that it has a consistent bias: it is too large. In Section 2.4.3 we described the conditions under which the Parzen density estimate is unbiased. When these conditions are met, a number of equalities hold:

$$
\begin{aligned}
p(X = x) &= \lim_{N_a \to \infty} P^*(x, a) && (3.32)\\
&= E_{\{a \in \{X\}\}}[P^*(x, a)] && (3.33)\\
&= E_{\{a \in \{X\}\}}\left[\frac{1}{N_a} \sum_{x_a \in a} R(x - x_a)\right] && (3.34)\\
&= E_X[R(X - x)] \ . && (3.35)
\end{aligned}
$$

Here $E_{\{a \in \{X\}\}}[P^*(x, a)]$ denotes the expectation over all possible random samples of size N_a drawn from the random variable X. Assuming the different samples of X are independent allows us to move the expectation inside the summation. The true entropy of the RV X can be expressed as

$$h(X) = -E_X[log\ E_{\{a \in \{X\}\}}[P^*(x, a)]] \ . \tag{3.36}$$

We can define a similar statistic:

$$\overline{h^*(X)} \equiv E_{\{b \in \{X\}, a \in \{X\}\}}[h^*(X)] \tag{3.37}$$

$$= -E_{\{b \in \{X\}, a \in \{X\}\}}[E_b[log E_a[R(x_b - x_a)]]] \tag{3.38}$$
$$= -E_X[E_{\{a \in \{X\}\}}[log(P^*(x, a))]] \ . \tag{3.39}$$

$\overline{h^*(X)}$ is the expected value of $h^*(X)$. Therefore, $h^*(X)$ provides an unbiased estimate of $\overline{h^*(X)}$. Jensen's inequality allows us to move the logarithm inside the expectation:

$$h(X) = -E_X[log E_{\{a \in \{X\}\}}[P^*(x, a)]] \tag{3.40}$$
$$\leq -E_X[E_{\{a \in \{X\}\}}[log(P^*(x, a))]] \tag{3.41}$$
$$\leq \overline{h^*(X)} \ . \tag{3.42}$$

The stochastic EMMA estimate is an unbiased estimator of a statistic that is provably larger than the true entropy. Intuitively, overly large estimates arise when elements of b fall in regions where $P^*(x, a)$ is too small. For these points the log of $P^*(x, a)$ is much smaller than it should be.

How then might we patch the definition of EMMA to remedy the bias? Another statistic that is similar to entropy is

$$\overline{\hat{h}(X)} = -E_X[p(X)] = -\int_{-\infty}^{\infty} p(x)^2 dx \ . \tag{3.43}$$

$\overline{\hat{h}(X)}$ is a measure of the randomness of X (we will define $\hat{h}(X)$ shortly). Strongly peaked distributions have large negative values for $\overline{\hat{h}}$. For widely spread uniform distributions $\overline{\hat{h}}$ approaches zero. Using the well known inequality, $x - 1 \geq log(x)$, we can show,

$$\overline{\hat{h}(X)} = -\int_{-\infty}^{\infty} p(x)^2 dx \tag{3.44}$$
$$\leq -\int_{-\infty}^{\infty} p(x) log(p(x)) dx \tag{3.45}$$
$$\leq h(X) \tag{3.46}$$

Parzen window density estimation can be used to construct a stochastic measure

$$\hat{h}(X) \equiv -E_b[P^*(X, a)] \ . \tag{3.47}$$

The expectation of $\hat{h}(X)$ is $\overline{\hat{h}(X)}$:

$$\begin{aligned} -E_{\{b\in\{X\},a\in\{X\}\}}[\hat{h}(X)] &= -E_{\{b\in\{X\},a\in\{X\}\}}[E_b[P^*(X,a)]] && (3.48)\\ &= -E_X[E_{\{a\in\{X\}\}}[P^*(X,a)]] && (3.49)\\ &= -E_X[p(X)] && (3.50)\\ &= \overline{\hat{h}(X)}\ . && (3.51)\end{aligned}$$

Further simplifying,

$$\overline{\hat{h}(X)} = -E_{x=X}[E_{\tilde{x}=X}[R(x-\tilde{x})]] \qquad (3.52)$$

is an expectation over a pair of events from the RV X. On average $\hat{h}$ is far too negative when $p(X)$ is large.

We have now defined two alternative statistics for which inexpensive unbiased estimates are available: $\overline{\hat{h}(X)}$ and $\overline{h^*(X)}$. These statistics bound the true entropy above and below,

$$\overline{\hat{h}(X)} \leq h(X) \leq \overline{h^*(X)}\ . \qquad (3.53)$$

h^* is on average too large and $\hat{h}$ is on average too small. Instead of using either, we have had good success using a third estimate,

$$\widetilde{h}(X) = -E_b[\Upsilon(x_b,a)]\ , \qquad (3.54)$$

where

$$\Upsilon(x,a) = \begin{cases} log(P^*(x,a)) & \text{if } P^*(x,a) < p_{min} \\ \frac{P^*(x,a)}{p_{min}} + log(p_{min}) - 1 & \text{otherwise} \end{cases}\ . \qquad (3.55)$$

The Υ function is designed so that both $\Upsilon(x,a)$ and $\frac{d\Upsilon(x,a)}{dx}$ are continuous. See Figure 3.1 for a plot of the Υ function. The intuition behind $\widetilde{h}$ is that we use h^* whenever possible, and for those sample points where $P^*(x,a)$ has a large standard deviation, we use $\hat{h}$ instead. The standard deviation of the Parzen density estimate is highest at points where the probability density is lowest, so we use $\hat{h}$ where $P^*(x,a)$ is below p_{min} (see Section 2.4.3). The variable p_{min} allows us to continuously vary the $\widetilde{h}$ estimate from the two extremes of $\hat{h}$ and h^*.

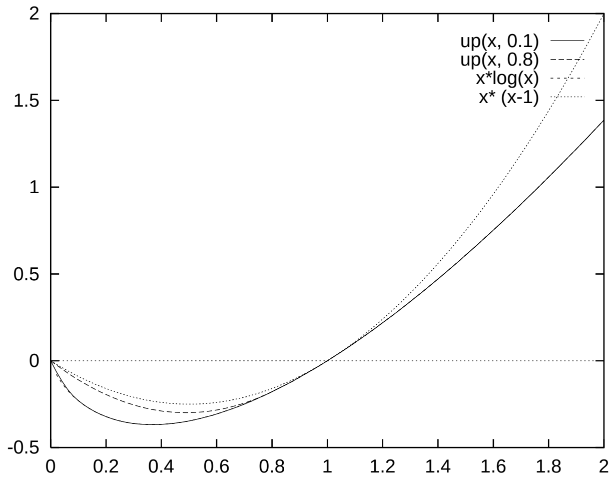


Figure 3.1: Plot of the functions $x\,\Upsilon(x)$ (which is labeled up), $x\,log(x)$, and $x\,(x-1)$. Two different values for p_{min} are plotted: 0.1 and 0.8. Notice that smaller values of p_{min} cause the approximation of $x\,log(x)$ to be very good. The difference between the two functions when $p_{min} = 0.1$ is almost unnoticeable.

Other Stochastic Search Techniques

Non-linear stochastic gradient descent is commonly used in the neural network literature, where it is often called the LMS rule. It was introduced there by Widrow and Hoff (Widrow and Hoff, 1960) and has been used extensively with good results. Since a stochastic estimate for the gradient of error is much cheaper to compute than a true estimate of the gradient, for many real problems LMS is faster than all other gradient techniques. The textbook by Haykin (Haykin, 1994) discusses the use of such algorithms by the neural network community. An excellent discussion of stochastic approximation appears in the textbook by Ljung and Söderström (Ljung and Söderström, 1983).

Simulated annealing is a related method that has been used in optimization problems which have many local minima (Kirkpatrick et al., 1983). These minima can "trap" gradient techniques far from the optimal solution. Simulated annealing performs a random, though usually local, search through parameter space. At each step a random modification to the parameters is proposed and the new cost is evaluated. If the new cost is lower than the previous cost the parameter modification is accepted. If the difference in cost is positive, the modification is accepted probabilistically. The probability of acceptance is proportional to the negative exponential of the difference,

$$p_{accept}(d) = \frac{exp(-\frac{d}{t})}{t} \quad , \tag{3.56}$$

where d is the difference between the new and the old cost and t is a temperature that controls the likelihood that a bad modification is accepted. Simulated annealing is based on the insight that physical systems, like iron, invariably find good energy minima when heated and then cooled slowly. The process of physical annealing is basically a gradient search perturbed by thermal noise. The thermal noise provides the energy to kick physical systems out of unfavorable local minima. The parameter t is an analog of physical temperature, it is initially set to large values and is gradually "cooled" during learning. In some cases it can be proven that with the right annealing schedule simulated annealing will converge to the global optimum.

Stochastic gradient descent can effectively penetrate *narrow* local minima which trap gradient techniques. Alignment applications can often have many narrow local minima which arise because of false matches between the high frequency components in the model and image. Because these false matches are based on features which are small the local minima are narrow in pose space. We have found that narrow local minima, which can trap gradient techniques, can be overcome using stochastic gradient descent.

We believe that stochastic approximation serendipitously combines efficient computation with effective escape from local minima.

3.3.1 Estimating the Covariance

In addition to the learning rate λ, the covariance matrices of the smoothing functions R are important parameters of EMMA. These parameters may be chosen so that they are optimal in the maximum likelihood sense. This is equivalent to minimizing the cross entropy of the estimated distribution with the true distribution (see Section 2.4.3). Our goal is to find the parameters that minimize empirical entropy.

For simplicity, we assume that the covariance matrices are diagonal,

$$\psi = \mathrm{DIAG}(\sigma_1^2, \sigma_2^2, \ldots) \ . \tag{3.57}$$

Following a derivation almost identical to the one described in Section 3.2 we can derive an equation analogous to (3.26),

$$\frac{d}{d\sigma_k} h^*(Y) = \frac{1}{N_b} \sum_{y_b \in b} \sum_{y_a \in a} W_y(y_b, y_a) \left(\frac{1}{\sigma_k}\right) \left(\frac{[y]_k^2}{\sigma_k^2} - 1\right) \tag{3.58}$$

where $[y]_k$ is the kth component of the vector y. This equation forms the basis for a method of stochastic maximization of likelihood.

Repeat:

$$A \leftarrow \{N_a \text{ points drawn from } y\}$$
$$B \leftarrow \{N_b \text{ points drawn from } y\}$$
$$\sigma_k \leftarrow \sigma_k + \lambda' \frac{d}{d\sigma_k} h^*(Y) \quad \forall k$$

The above procedure is very similar to the one described in Section 3.3. During entropy manipulation, it is possible to interleave covariance updates with parameter updates.

3.4 Principal Components Analysis and Information

As a demonstration, we can derive a parameter estimation rule akin to principal components analysis that truly maximizes information. This new EMMA based component analysis (ECA) manipulates the entropy of the random variable $Y_v = X \cdot v$ under the constraint that $|v| = 1$. For any given value of v the entropy of Y_v can be estimated from a sample of X as:

$$h^*(Y_v) = \frac{1}{N_b} \sum_{y_b \in b} log \left(\frac{1}{N_a} \sum_{y_a \in a} g_\psi(y_b - y_a) \right) \tag{3.59}$$

$$= \frac{1}{N_b} \sum_{x_b \in b} log \left(\frac{1}{N_a} \sum_{x_a \in a} g_\psi(x_b \cdot v - x_a \cdot v) \right) \tag{3.60}$$

where ψ is the variance of the Parzen smoothing function. Moreover we can estimate the derivative of entropy:

$$\frac{d}{dv}h^*(Y_v) = \frac{1}{N_b}\sum_{y_b\in b}\sum_{y_a\in a} W_y(y_b, y_a)\ \frac{1}{\psi}(y_b - y_a)\frac{d}{dv}(y_b - y_a) \tag{3.61}$$

$$= \frac{1}{N_b}\sum_b\sum_a W_y(y_b, y_a)\ \frac{1}{\psi}(y_b - y_a)(x_b - x_a)\ \ . \tag{3.62}$$

Let us decompose the derivative into parts which can be understood more easily. We will first analyze the second part of the summand: $(y_b - y_a)(x_b - x_a)$. Ignoring the weighting function $W_y\psi^{-1}$ we are left with the derivative of some unknown function $f(Y_v)$:

$$\frac{d}{dv}f(Y_v) = \sum_b\sum_a(y_b - y_a)(x_b - x_a) \tag{3.63}$$

$$= N_b N_a E_b[E_a[(y_b - y_a)(x_b - x_a)]]\ \ . \tag{3.64}$$

What then is $f(Y_v)$? The derivative of the squared difference between samples is:

$$\frac{d}{dv}(y_b - y_a)^2 = 2(y_b - y_a)\frac{d}{dv}(x_b\cdot v - x_a\cdot v) \tag{3.65}$$

$$= 2(y_b - y_a)\frac{d}{dv}[(x_b - x_a)\cdot v] \tag{3.66}$$

$$= 2(y_b - y_a)(x_b - x_a)\ \ . \tag{3.67}$$

So we can see that

$$f(Y_v) = N_b N_a E_b[E_a[(y_b - y_a)^2]] \tag{3.68}$$

is the expectation of the squared difference between pairs of trials of Y_v.

Recall that PCA searches for the RV Y_v that has the largest variance: $E_a[(y_a - E_a[y_a])^2] = Var_a(Y_v)$. Interestingly the expected squared difference between a pair of trials is precisely twice the variance:

$$C(v) = E_b[E_a[(y_b - y_a)^2]] \tag{3.69}$$

$$= E_b[E_a[y_b^2 - 2y_a y_b + y_a^2]] \tag{3.70}$$

$$= E_b[y_b^2 + E_a[2y_a y_b + y_a^2]] \tag{3.71}$$

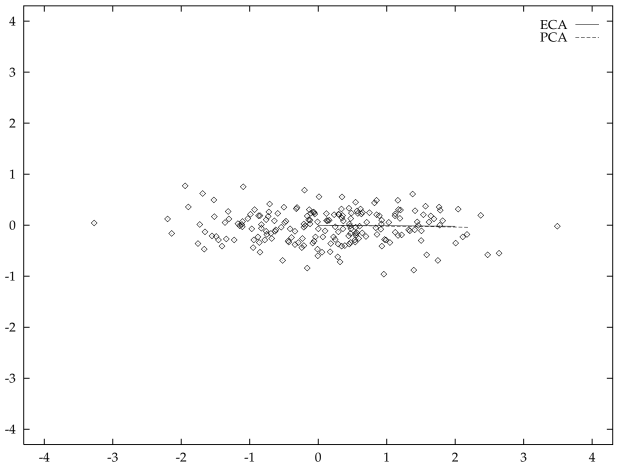


Figure 3.2: A scatter plot of a sample from a two dimensional Gaussian density. The sample contains 200 points. The principal axis and the ECA axis are also plotted as vectors from the origin. The vectors are nearly identical.

$$= E_b[y_b^2 - 2y_b E_a[y_a] + E_a[y_a^2]] \tag{3.72}$$

$$= E_b[y_b^2] - 2E_b[y_b]E_a[y_a] + E_a[y_a^2]] \tag{3.73}$$

$$= E_b[Y^2] - 2E_b[Y]E_a[Y] + E_a[Y^2] \tag{3.74}$$

$$= 2Var_a(Y) \ . \tag{3.75}$$

Without the weighting term, $W_y\psi^{-1}$, ECA would find exactly the same vector that PCA does: the maximum variance projection vector. However the derivative of ECA does not act on all points of Y_v equally. Recall that $W_y(y_a, y_b)$ is a measure of the distance between y_b and y_a. It is large when y_b is significantly closer to y_a than any other element of a. As a result ECA maximizes variance in a local way. Points that are very far apart are forced no further apart. Another way of interpreting (ECA) is as a type of robust variance maximization. Points that might best be interpreted as outliers, because they are very far from the body of other points, play a very small role in the minimization. These robust characteristic stand in contrast to PCA which is very sensitive to outliers.

For densities that are Gaussian, the maximum entropy projection is the first principal component. In simulations ECA effectively finds the same projection as PCA. Figure 3.2 shows a sample of data and the PCA and ECA principal components. Since this density has a larger variance along the horizontal axis, both the ECA and PCA axes point along the horizontal axis. Our ECA code take roughly 10 seconds to run on Sun Sparc5 workstation.

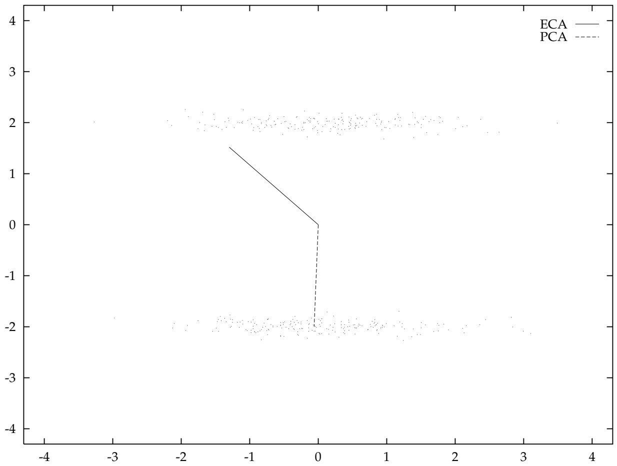


Figure 3.3: A scatter plot of a 400 point sample from a two dimensional density. The density is a mixture of two horizontally stretched Gaussians. The PCA and ECA principal axes are also plotted as vectors from the origin.

This is comparable to the time it takes to run PCA.

In general, PCA does not find the highest entropy projection of non-Gaussian densities. For more complex densities the PCA axis is very different from the entropy maximizing axis. Figure 3.3 shows a density for which the PCA and ECA axes are very different. The PCA axis, which is vertical, spreads the points in the sample as far apart as possible. The ECA axis, which is oblique, spreads nearby points in the sample as far apart as possible. The resulting densities, Y_{PCA} and Y_{ECA}, are graphed in Figure 3.4. The PCA density is very tightly peaked, the ECA density is broadly spread out. Though the final variance of Y_{PCA} is larger, 2.005 vs. 1.626, the entropy of the Y_{ECA} distribution is much higher, $h^*(Y_{PCA}) = -0.17$ and $h^*(Y_{ECA}) = 1.61$.

Linsker has argued that the PCA axis separates the clusters of a distribution (Linsker, 1988). To justify this claim, he uses figures much like Figure 3.3 and Figure 3.4. These graphs show the PCA axis projecting points from separated clusters so that they remain separate. It is then proposed that the PCA axis is useful for cluster classification of high dimensional data. In other words, that high dimensional data can be projected down into a low dimensional space without perturbing the cluster structure. In general this is *not* true. PCA only separates clusters when the variance between clusters is higher than the variance within clusters.

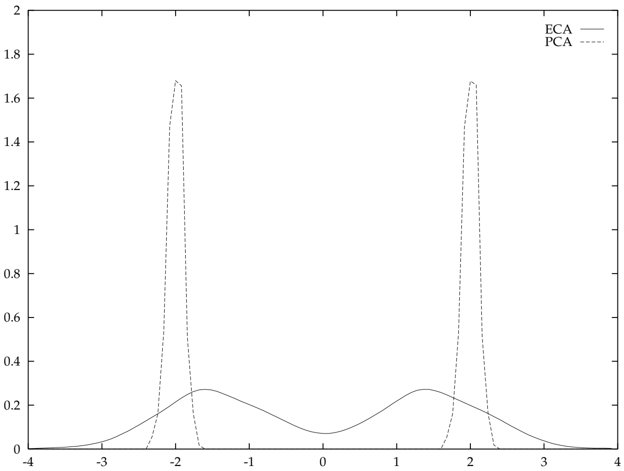


Figure 3.4: The Parzen density estimates of Y_{PCA} and Y_{ECA}.

Ironically, it is the *minimum* entropy projection that should separate clusters well. Let us assume that each cluster is generated from a prototypical point that has been perturbed by random noise. If there is very little noise, the sample points associated with a cluster prototype will be clustered together tightly. The resulting density is sharply peaked around the cluster prototypes and has low entropy. Additional noise acts to spread out each cluster, adding entropy to the density. Most of the entropy in this density arises from the noise, not the clusters. An entropy maximizing algorithm will find a projection vector that maximizes the projection of the noise. On the other hand, an entropy minimizing algorithm should, if possible, find a projection that is perpendicular to the noise. ECA can be used both to find the entropy maximizing (ECA-MAX) and minimizing (ECA-MIN) axes.

Figure 3.5 shows a distribution where the noise, or spread, of the clusters is perpendicular to the axis that separates the clusters. As a result, the PCA axis does not separate these clusters. The ECA axis shown is the minimum entropy axis (which is obtained by running the EMMA algorithm with a negative learning rate). The ECA-MIN axis separates the clusters much better than the PCA axis (see Figure 3.6).

To provide further intuition regarding the behavior of ECA we have run ECA, PCA, and two related procedures BCM and BINGO on the same density. BCM is a learning rule that was originally proposed to explain development of receptive fields patterns in visual cortex (Bienenstock et al., 1982). More recently it has been argued that the rule finds projections that are far from Gaussian (Intrator and Cooper, 1992). Under a limited set of conditions

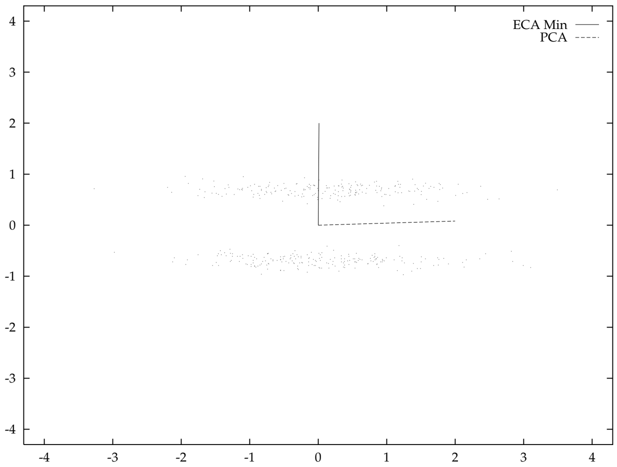


Figure 3.5: A scatter plot of a 400 point sample from a two dimensional density. The density is a mixture of two horizontally stretched Gaussians. The PCA and ECA minimum entropy axes are also plotted as vectors from the origin.

BCM finds the minimum entropy projection. BINGO was proposed to find axes along which there is a bimodal distribution (Schraudolph and Sejnowski, 1993).

Figure 3.7 displays a 400 point sample and the five different projection axes found by the algorithms discussed above discussed above. The density is a mixture of two clusters. Each cluster has high kurtosis in the horizontal direction. The oblique axis projects the data so that it is most uniform and hence has the highest entropy; ECA-MAX finds this axis. Along the vertical axis the data is clustered and has low entropy; ECA-MIN finds this axis. Interestingly because the vertical axis has high variance, PCA finds the entropy minimizing axis. BCM, while it may find minimum entropy projections for some densities, is attracted to the kurtosis along the horizontal axis. The horizontal axis neither minimizes nor maximizes entropy. Finally, BINGO successfully discovers that the vertical axis is very bimodal. The densities of the different projections is shown in Figure 3.8.

3.5 Conclusion

This chapter has presented a new technique for estimating the entropy of a distribution called EMMA. Provided the density being approximated is smooth, we have proven that the technique will converge to the correct entropy estimate. Moreover we have presented

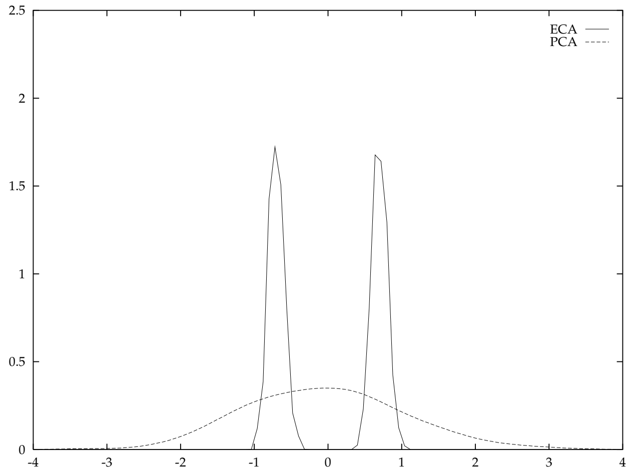


Figure 3.6: The Parzen density estimates of Y_{PCA} and Y_{ECA} from the previous graph.

a computationally efficient stochastic technique for manipulating entropy. For reasonable sample sizes, the technique is not guaranteed to optimize true entropy. Instead it optimizes a very similar statistic that retains all of the salient characteristics of entropy.

We have also described a simple application of EMMA. EMMA enables us to find low dimensional projections of higher dimensional data that minimize or maximize entropy.

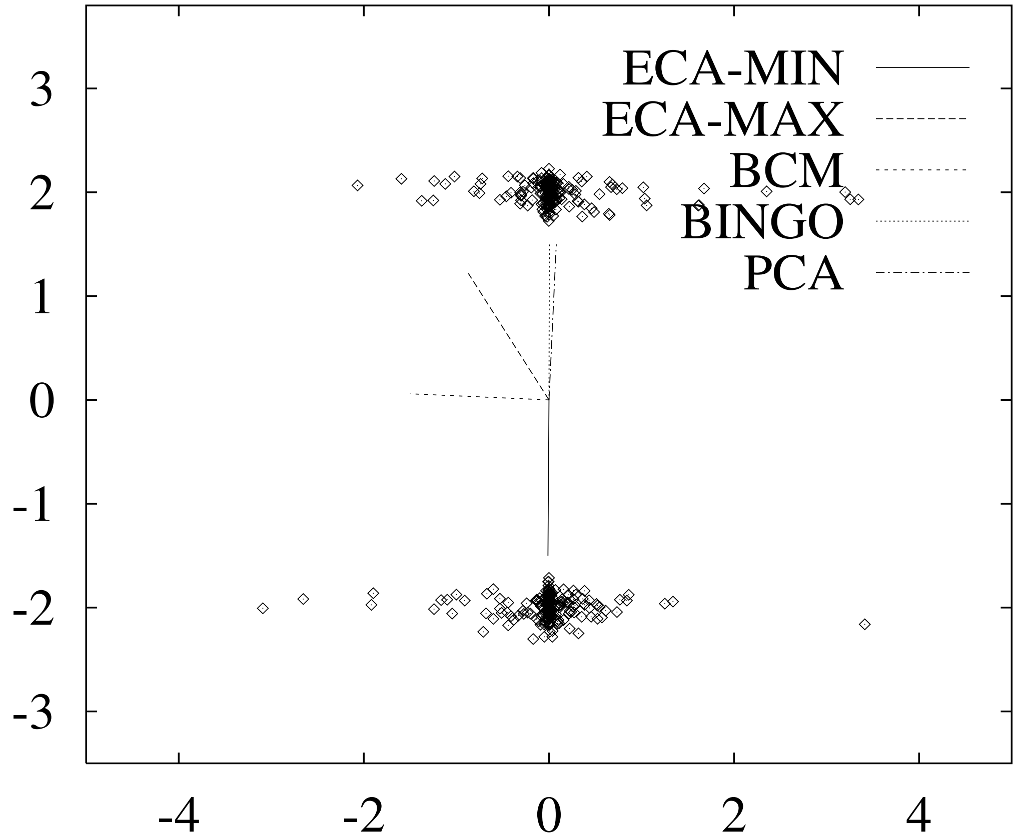


Figure 3.7: A scatter plot of a 400 point sample from a two dimensional density. Each cluster has very high kurtosis along the horizontal axis. See text for description of projection axes.

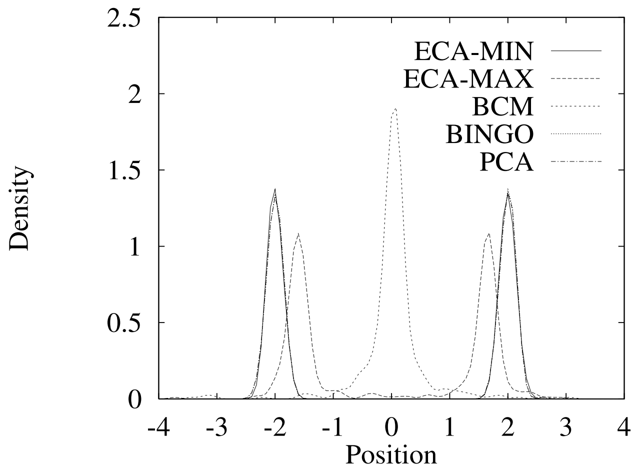


Figure 3.8: The densities along various projection axes.

Chapter 4

Matching and Alignment

This chapter is perhaps the most important in this thesis. Previous chapters have presented the mathematics and algorithms that underly the computation of empirical entropy. We have already seen that empirical entropy can be used to define a new algorithm for finding the most informative projection of a distribution. This chapter will show that matching and alignment can also be formulated as an entropy problem. In addition, we will discuss the intuition behind our framework and suggest some simplified schemes that reflect these intuitions. Throughout this chapter a number of synthetic alignment problems will drive our discussions.

We will begin with a rederivation of correlation as a maximum likelihood method. This derivation will make clear the assumptions under which correlation will work, and when it may fail. We will then attempt to generalize correlation so that it will work with a wider set of inputs. While this generalization is theoretically straightforward it will prove computationally intractable.

Dropping our focus on correlation, we will define an intuitive approach to alignment which is efficiently computable. Using this intuition we will then define an approximation to a maximum likelihood technique that is both concrete and computable. Finally we will draw a parallel between this technique and mutual information. Experimental data from synthetic alignment problems will help us evaluate the proposed alignment techniques.

This chapter will conclude with an entirely different motivation for the use of mutual information in alignment. We will show how the alignment problem can be thought of as a *Minimum Description Length* problem (Rissanen, 1978; Leclerc, 1988). This formulation will naturally focus on the task of coding efficiency and entropy minimization. A very similar set of alignment equations will arise from these considerations.

4.1 Alignment

We are given two signals of time or space: $u(x)$ and $v(y)$. We will call $u(x)$ the *model*. Often it is a description of a physical object that has been computed with great care. For example, in one of our experiments the model is an accurate three dimensional description of a skull. The second signal, $v(y)$, is an *image* of the model. In general the form and even the coordinate systems of the model and image can be very different. For example, one of our 3D models is a collection of three dimensional points and normals; the corresponding image is a two dimensional array of intensities. It is assumed that $v(y)$ is an observation of $u(x)$, for example that $v(y)$ is a picture of the skull $u(x)$.

The relationship between $u(x)$ and $v(y)$ is based on the physics of imaging. The process of constructing an observation has two separate components. The first component is called a transformation, or pose, denoted T. It relates the coordinate frame of the model, x, to the coordinate frame of the image, y. The transformation tells us which part of the model is responsible for a particular pixel of the image. The second component is the imaging function, $F(u(x), q)$. The imaging function determines the value of image point $v(T(x))$. In general a pixel's value may be a function both of the model and other exogenous factors. For example, an image of an object depends not only on the object but also on the lighting. The parameter, q, collects all of the exogenous influences into a single vector. The complete imaging model is then:

$$v(T(x)) = F(u(x), q) + \eta \ , \tag{4.1}$$

or equivalently,

$$v(y) = F(u(T^{-1}(y)), q) + \eta \ , \tag{4.2}$$

where η is a random variable that models noise in the imaging process.

For a number of practical problems, the transformation between a model and an image

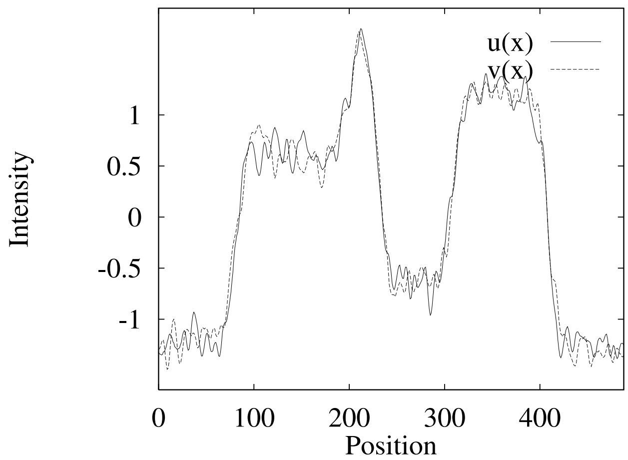


Figure 4.1: Graph of $u(x)$ and $v(x) = u(x) + \eta$ versus x.

is not known. Alignment is the process by which the correct transformation is extracted. Alignment can be a difficult problem for a number of reasons:

- The imaging function F of the physical world can be difficult to model.
- The exogenous parameters q are not necessarily known and can be difficult to find. For example computing the lighting in an image is a non-trivial problem.
- The space of transformations, which may have many dimensions, is difficult to search. Rigid objects often have a 6 dimensional transformation space. Non-rigid objects can in principle have an unbounded number of pose parameters.

A simple example can lend intuition to these definitions. Let $u(x)$ and $v(y)$ be one dimensional signals. Let the transformation space be the space of all possible translations

$$T(x) = x - \beta \quad . \tag{4.3}$$

Let the imaging function F be the identity function. Choosing $\beta = 0$ leads to

$$v(x) = u(x) + \eta \quad . \tag{4.4}$$

Figure 4.1 contains a graph of two signals that obey this relationship. Though we show the image and model aligned, the correct alignment between v to u may not be known.

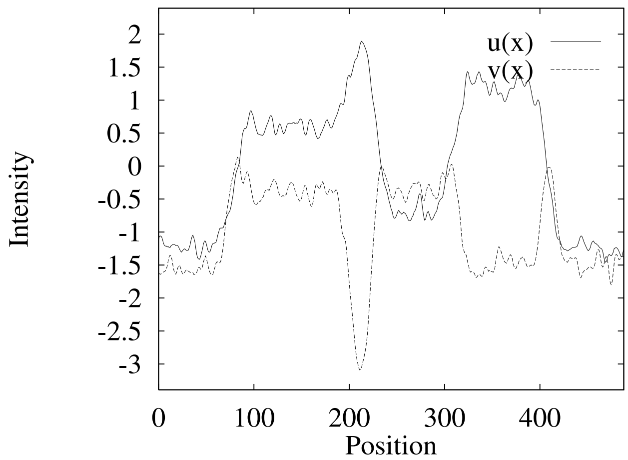


Figure 4.2: Graph of $u(x)$ and $v(x) = -u(x)^2$ versus x.

In all of our synthetic experiments 10% random noise has been added to v[1]. Noise is of course an unavoidable reality of any real system. But more importantly the addition of noise demonstrates that the algorithms presented are numerically stable.

More complex imaging functions are possible. For example, F might be non-linear

$$F(u) = -u^2 \quad . \tag{4.5}$$

Figure 4.2 contains a graph of $u(x)$ and $v(T(x)) = F(u(x))$.

4.1.1 Correlation as a Maximum Likelihood Technique

The search for the correct alignment can be cast as a maximum likelihood or variance minimization problem (see Section 2.3.1). The probability of an image given the model, the transformation, the noise distribution, the exogenous parameters, and the imaging function is:

$$p(v \mid u, T, \eta, q, F) = \prod_{x_a \in a} p\left(\eta = (v(T(x_a)) - F(u(x_a), q))\right) \quad . \tag{4.6}$$

[1]We use white noise that has been low-passed filtered to roughly 0.3 cycles per unit. The peak to peak amplitude of the noise is 10% of the peak to peak amplitude of the signal.

In the above equation we have assumed that each pixel of v is conditionally independent. Conditional independence does not imply that the pixels are independent, just that if u, T, η, q, and F are known the pixels are independent. Assuming that the noise is Gaussian, we can then compute the log likelihood of a transformation as

$$\begin{aligned}
log(\ell(T)) &= log\ p(v \mid u, T, \eta, q, F) && (4.7)\\
&= \sum_{x_a \in a} log\ p\left(\eta = v(T(x_a)) - F(u(x_a), q)\right) && (4.8)\\
&= -k_1 \sum_{x_a \in a} \left(v(T(x_a)) - F(u(x_a), q)\right)^2 && (4.9)\\
&\approx -k_2 E\left[\left(v(T(X)) - F(u(X), q)\right)^2\right] && (4.10)\\
&\approx -k_2 E\left[v(T(X))^2\right] - 2E\left[v(T(X))F(u(X), q)\right] + E\left[F(u(X), q)^2\right] && (4.11)
\end{aligned}$$

where k_1 and k_2 are constants computed from the variance of the noise and the number of sample points. They play no role in the maximization. In (4.11) we have expanded the square of the difference to show that the log likelihood of a transformation has three components: one that arises from the variance of the model; a second that arises from the correlation between the image and the predicted image; and a third that arises from the variance of the predicted image. For problems where the variance of the image and predicted image are fixed, the best transformation is the one that maximizes the correlation between the actual and predicted image.

For convenience we will define the cost of a transformation as

$$\begin{aligned}
C(T) &= E\left[\left(v(T(X)) - F(u(X), q)\right)^2\right] && (4.12)\\
&\propto -log(\ell(T))\ . && (4.13)
\end{aligned}$$

The lowest cost transformation is the one that causes the model to match the image "best".

As we did in the analysis of principal components and function learning, we have "invented" random variables: X, $v(T(X))$ and $u(X)$. The random variable X ranges over points from the coordinate system of u where $u(X)$ is defined. The random variables $u(X)$ and $v(T(X))$ range over the values in the model and image. In reality there are no random processes involved in matching and alignment. The model and image are pre-determined and fixed. Alignment could proceed deterministically; the cost of a transformation being evaluated directly from all of the points in the model and image. We have chosen to interpret the summation over pixels that arises in correlation as an expectation over a set of random variables. As a result the insights of probability and statistics can be brought to bear on these problems.

4.1.2 Correlation and Mutual Information

Alignment is very similar to the problem of function learning that we encountered in Section 3.1. Equation (3.15) is almost identical to (4.7). In both problems we are looking for a set of parameters that cause the inputs to model the outputs. Function learning attempts to find the best parameter vector v; alignment attempts to find the best transformation T. As we did for function learning, we can draw an analogy with sample entropy. The log likelihood of T is proportional to the conditional entropy of the image given the model,

$$log(\ell(T)) = -\frac{1}{N_a} h_a(v(T(X)) \mid u(X), T, \eta, q, F) \ . \tag{4.14}$$

This is not the EMMA estimate of entropy, but the conditional entropy of v under the assumption that v is conditionally Gaussian. For the problems described in Section 3.1 it was possible to show that entropy optimization led to maximum mutual information solutions. For this problem however, we cannot claim that maximizing log likelihood is equivalent to maximizing the mutual information between v and u. The mutual information

$$I(v(T(X)), u(X)) = h(v(T(X))) - h(v(T(X)) \mid u(X), T, \eta, q, F) \ , \tag{4.15}$$

includes both a conditioned and unconditioned entropy. For some types of transformations $h(v(T(X)))$ may change as T is varied. In these cases minimizing conditional entropy is not equivalent to maximizing mutual information. One must also maximize the unconditioned entropy. In our simple example, where only translation is varied and the signals are periodic, unconditioned entropy does not change as T is varied.

Returning to the first synthetic example, we can plot $C(T)$ (from (4.12)) versus translation. Figure 4.3 graphs $C(T)$ for the two signals from Figure 4.1 (we have assumed periodic boundary conditions on the signals). We can see that the cost has a very strong minimum at the true translation of 0 pixels.

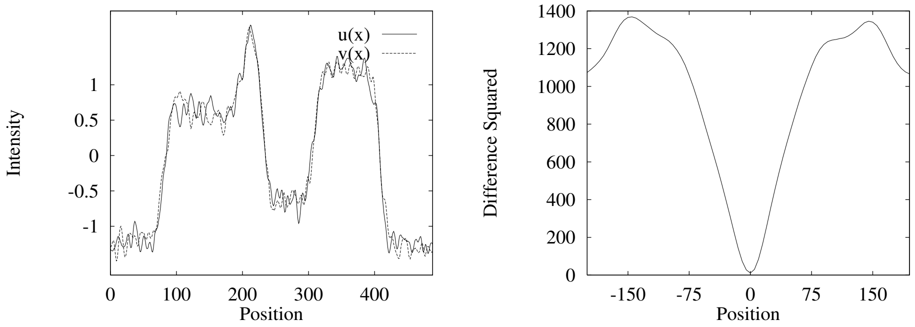


Figure 4.3: On the left is a plot of image and model that are identical except for noise. On the right is a plot of $C(T)$ versus translation. There is a significant minimum at the correct aligning translation of 0 pixels.

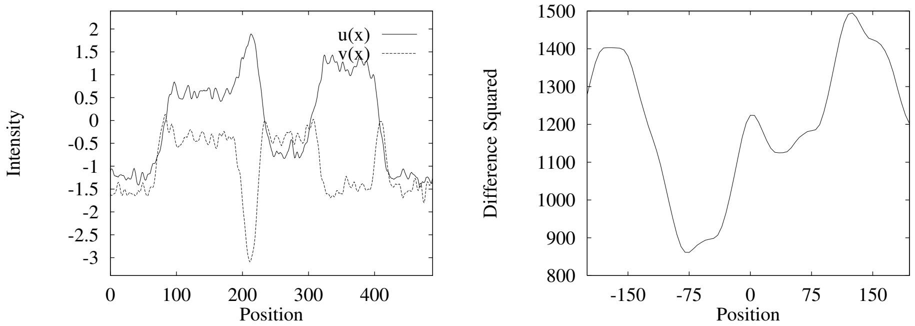


Figure 4.4: On the left is a plot of image and model that are related non-linearly. On the right is a plot of $C(T)$ versus translation. There is no minima at the aligning translation of 0 pixels. In fact minima exist at incorrect translations.

Correlation works very well at matching together u and v when the imaging model and exogenous parameters are known. In many cases however we may be faced with a situation where F and q are *unknown*. In some cases alignment problems can still be solved by assuming that the imaging function is the identity function. This assumption is not effective when aligning the non-monotonically related signals shown in Figure 4.2. Figure 4.4 graphs $C(T)$ versus translation for these two signals. Notice that each of the actual minima are at incorrect translations.

In general $C(T)$ cannot be used to align signals related by an unknown non-linear transformations. $C(T)$ can however be generalized to work with signals that have been transformed linearly. Rather than minimize the squared difference between signals, we can instead minimize the squared difference between signals that have been normalized. A normalized signal

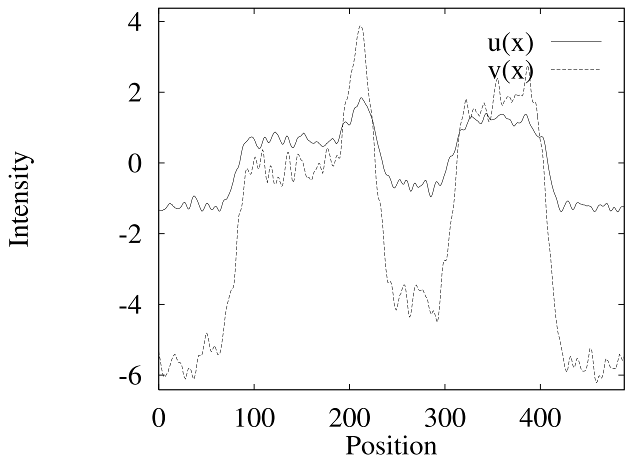


Figure 4.5: Graph of $u(x)$ and $v(x) = 3u(x) - 2$ versus x.

is one with a mean of zero and a standard deviation of one and can be computed as

$$\widehat{u}(x) = \frac{u(x) - E[u(X)]}{\sigma(u(X))} \quad . \tag{4.16}$$

The normalized version of a signal is invariant to multiplicative and additive changes to the original. The sum of the squared differences between the normalized signals, $NC(T)$, can be computed directly as one minus the *Normalized correlation* between the signals u and v. Normalized cost is defined as:

$$NC(T) = 1 - \frac{E_a[u(X)v(T(X))] - E_a[u(X)]E_a[v(T(X))]}{\sigma_a(u(X))\sigma_a(v(T(X)))} \quad . \tag{4.17}$$

As a shorthand we have abbreviated sums over the coordinates x as expectations and variances.

Normalized cost can be used on signals like the ones shown in Figure 4.5 ($F(u) = 3u - 2$). A plot of $NC(T)$ versus translation is identical to Figure 4.3. In some cases, normalized cost can be applied to signals transformed by non-linear *monotonic* functions. Note however that the two signals shown in Figure 4.2 are related by a non-monotonic function and cannot be accommodated in this way. In these examples translation does not effect the mean or the standard deviation of the signals. As a result, normalized cost will not produce a convincing minimum where cost alone does not.

We may still wish to align models and images that are related by non-monotonic functions. In this case alignment can be performed by jointly searching over the space of possible imaging functions, exogenous parameters and transformations. Probability can be used to motivate this procedure. We can evaluate $p(v \mid u, T, N, q, F)$ when F and q are unknown by integrating out the unknown variables. The probability of the image would then be,

$$p(v \mid u, T, \eta) = \int \int \prod_{x_a \in a} p(\eta = v(T(x_a)) - F(u(x_a), q)) \; p(F) \; p(q) \; dF \; dq \quad . \tag{4.18}$$

This equation integrates over all possible imaging functions and all possible sets of exogenous variables. We are not aware of any approach that has come close to evaluating such an integral. It may not be feasible. Another possible approach is to find the imaging function and exogenous variables that make the image most likely,

$$p(v \mid u, T, N) \approx \max_{F,q} \prod_{x_a \in a} p(\eta = v(T(x_a)) - F(u(x_a), q)) \; p(F) \; p(q) \quad . \tag{4.19}$$

Here we have assumed that the integral in Equation 4.18 is approximated by the component of the integrand that is maximal. The approximation is a good one when a particular F and q are much more likely than any other.

Using (4.19) we can define an alignment procedure as a nested search: (1) given an estimate for the transformation, find F and q that make the image most likely; (2) given estimates for F and q, find a new transformation that makes the image most likely. Terminate when the transformation has stabilized. In other words, a transformation associates points from the model with points in the image; for every $u(x)$ there is a corresponding $v(T(x))$. A function F and parameter vector q are sought that best model the relationship between $u(x)$ and $v(T(x))$. This can be accomplished by "training" a function to fit the collection of pairs $\{v(T(x_a)), u(x_a)\}$. Algorithms for finding F and q are very similar to the those for density approximation and learning described in Chapter 3. Notice also that that alignment with an unknown imaging model is very similar to entropy maximization. Entropy maximization is a nested search for a density estimate and parameters. Alignment is a nested search for an imaging model and a transformation. We will return to this analogy shortly.

Many of the pitfalls of density approximation as described in Chapter 2 apply to function approximation as well. Before we can hope to learn the function F we must first make a set of assumptions about the form of F. Without these assumptions discontinuous estimates for

F, which fit the data perfectly well but are very unlikely, can prevent convergence. One way to prevent, or discourage, this behavior is to formulate a strong prior probability over the space of functions, $p(F)$.

In many cases the search for an imaging function and exogenous parameters can be combined. For any particular F and q, another function $F_q(u(x)) = F(u(x), q)$ can be defined. Combining functions like this is a common technique in both "shape from shading" and "photometric stereo" research. Both techniques compute the shape of an object from the shading that is present in an image or images. Rather than independently model the exogenous variable (the lighting direction) and imaging function (the reflectance function) a combined function is represented and manipulated. The combined function is called a *reflectance map* (Horn, 1986). It maps the normals of an object directly into intensities. The three dimensional alignment procedure we will describe manipulates a similar combined function.

How might Equation 4.19 be approximated efficiently? It seems reasonable to assume that for most real imaging functions similar inputs should yield similar outputs. In other words, that the unknown imaging function is continuous and piecewise smooth. An efficient scheme for alignment could skip the step of approximating the imaging function and attempt to directly evaluate the *consistency* of a transformation. A transformation is considered consistent if points that have similar values in the model project to similar values in the image. By similar we do not mean similar in physical location, as in $|x_a - x_b|$, but similar in value, $|u(x_a) - u(x_b)|$ and $|v(T(x_a)) - v(T(x_b))|$. One ad-hoc technique for estimating consistency is to pick a similarity constant ϵ and evaluate the following sum:

$$\text{Consistency}_1(T) = -\sum (v(T(x_b)) - v(T(x_a)))^2 \ \ , \tag{4.20}$$

where the sum is over $x_a \in a$ and $x_b \in b$ such that $|u(x_b) - u(x_a)| < \epsilon$ and $x_a \neq x_b$. Consistency is flawed in a number of ways. For instance, there are no obvious clues of picking ϵ. We can replace the "all or nothing" nature of the ϵ test with a more gradual discrimination:

$$\text{Consistency}_2(T) = -\sum_{x_a \neq x_b} g_\psi(u(x_b) - u(x_a))(v(T(x_b)) - v(T(x_a)))^2 \ \ , \tag{4.21}$$

where g_ψ is a Gaussian with standard deviation, ψ. In order to minimize this measure, points that are close together must be more consistent, and those further apart less so. Another problem with any consistency measure is that it is too aggressive; consistency is maximized by constancy. The most consistent transformation projects the points of the model onto a constant region of the image. For example, if scale is one of the transformation parameters, one entirely consistent transformation projects all of the points of the model down to a single point of the image. Though there are a number of problems with consistency that need to be addressed, it will serve as source of intuition as we analyze different approaches.

We now have two alternatives for alignment when the imaging function is unknown: a theoretical technique that may be intractable (4.19), and an outwardly efficient technique that has a number of important difficulties (4.21). One would like to find a technique that combines the best feature from each approach. Perhaps the complex search for the most likely imaging function, F_q, can be replaced with a simpler search for a consistent imaging function.

One type of function approximator that maximizes consistency is known as a nearest neighbor function approximator (Duda and Hart, 1973). A nearest neighbor function approximator $F_N(u, a)$ is constructed directly from a sample a. The approximator's value at a point u is the value of the point which is nearest in the sample:

$$F_N(u, a) = v(T(\hat{x})) \text{ such that } \hat{x} = \underset{x_a \in a}{\operatorname{argmin}} |u - u(x_a)| \ . \tag{4.22}$$

F_N can be used to estimate the likelihood of a model as we did in (4.19). The nearest neighbor formulation can be much more efficient than a naive implementation of (4.19), since there is no need to search for F_q. The model, image, and transformation define F_N directly.

The nearest neighbor function approximator plays a role in the likelihood computation that is very similar to the role that Parzen density estimation plays in entropy estimation (see Sections 2.4.3 and 3.2). Unlike the Parzen estimate, the nearest neighbor approximation is not continuously differentiable. A similar though differentiable version is called a *weighted neighbor* approximator[2]:

$$F^*(u, a) = \frac{\sum_a R(u - u(x_a))v(T(x_a))}{\sum_a R(u - u(x_a))} \ . \tag{4.23}$$

The weighting function R usually has a maximum at zero, and falls off asymptotically away

[2]This technique is also known as kernel regression.

from zero. A common choice for R is the Gaussian density function g_ψ with standard deviation ψ. We can re-write F^* as,

$$F^*(u,a) = \sum_a W_a(u, u(x_a)) v(T(x_a)) \ . \tag{4.24}$$

where W is the soft nearest neighbor function first defined in Section 3.2,

$$W_a(u_1, u_2) \equiv \frac{g_\psi(u_1 - u_2)}{\sum_a g_\psi(u_1 - u(x_a))} \ .$$

An estimate for the log likelihood is then,

$$\begin{aligned}
log(\ell(T)) &= -k_2 \sum_a [v(T(x_a)) - F^*(u(x_a), a)]^2 && (4.25)\\
&= -k_2 \sum_a \left[v(T(x_a)) - \sum_b W_a(u(x_a), u(x_b)) v(T(x_b)) \right]^2 && (4.26)\\
&= -k_2 \sum_a \left[\sum_b W_a(u(x_a), u(x_b)) v(T(x_a)) - \sum_b W_a(u(x_a), u(x_b)) v(T(x_b)) \right]^2 && (4.27)\\
&= -k_2 \sum_a \left[\sum_b W_a(u(x_a), u(x_b)) (v(T(x_a)) - v(T(x_b))) \right]^2 \ . && (4.28)
\end{aligned}$$

Step (4.27) relies on the fact that

$$\sum_b W_a(u(x_a), u(x_b)) = 1 \ ,$$

where $x_a \in a$. The log likelihood of a transformation using a weighted neighbor function approximation is very similar to the intuitive consistency measure (4.21). In addition its derivative bears a striking resemblance to the derivative of the EMMA estimate (see (3.26)):

$$\begin{aligned}
&\frac{d}{dT} log(\ell(T)) && (4.29)\\
&= -k_2 \sum_a \frac{\sum_b \left(g_\psi(u(x_a) - u(x_b))^2 (v(T(x_a)) - v(T(x_b))) (\frac{d}{dT} v(T(x_a)) - \frac{d}{dT} v(T(x_b))) \right)}{\left(\sum_b g_\psi(u(x_a) - u(x_b)) \right)^2} && (4.30)\\
&= -k_2 \sum_a \sum_b W_a(u(x_a), u(x_b))^2 \left((v(T(x_a)) - v(T(x_b))) \frac{d}{dT} (v(T(x_a)) - v(T(x_b))) \right) && (4.31)
\end{aligned}$$

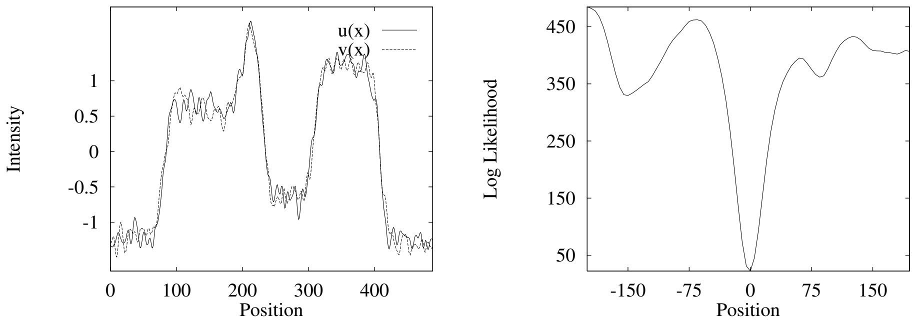


Figure 4.6: On the left is a plot of image and model that are identical except for noise. On the right is a plot of the logarithm of weighted neighbor likelihood versus translation.

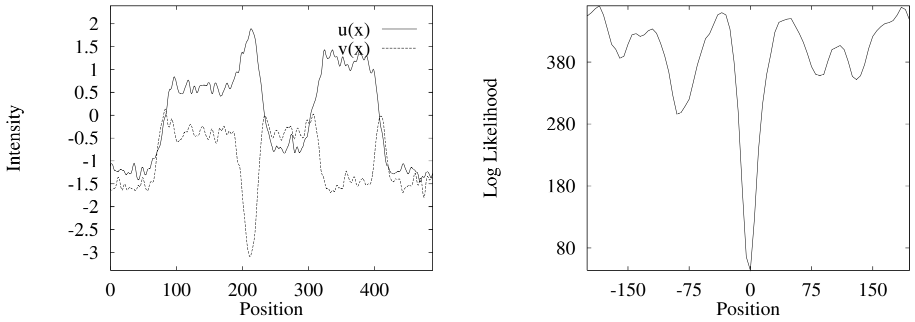


Figure 4.7: On the left is a plot of image and model that are related non-linearly. On the right is a plot of the logarithm of weighted neighbor likelihood versus translation.

Weighted neighbor likelihood can be used to evaluate the cost of different translations. Figure 4.6 shows a graph of weighted neighbor likelihood versus translation for the initial pair of signals, $u(x)$ and $v(x) = u(x) + \eta$. Figure 4.7 contains a similar graph for the second, non-linear experiment, $u(x)$ and $v(x) = (u(x) - 2)^2 + \eta$. Both graphs show a strong minimum at the correct alignment of the signals. We can conclude that weighted neighbor likelihood can be used in situations where neither cost nor normalized cost would work.

The parallel between EMMA and weighted neighbor likelihood is more than structural. EMMA estimates the density of the sample directly and uses it to compute the derivative of entropy with respect to the parameter vector; weighted neighbor likelihood estimates the imaging function directly and uses it to compute the derivative of log likelihood with respect to the transformation. More importantly both techniques manipulate the entropy of the of the joint distribution u and v. EMMA can be used to evaluate the joint entropy $h^*(v(T(x)), u(X))$; weighted neighbor likelihood evaluates the conditional entropy $h(v(T(x)) \mid u(X))$ under the assumption that $p(v(T(x))|u(X), a) = g_\psi(v(T(x)) - F^*(u(x), a))$ (see Sections 2.3.1 and 4.1.2 for commentary on the equivalence of log likelihood and sample entropy).

We can relax the constraint that v be conditionally Gaussian by using EMMA to estimate the conditional entropy:

$$h(v(T(x))|u(X)) \approx h^*(u(x)) - h^*(v(T(x)), u(X)) \ . \tag{4.32}$$

The first term is the entropy of the model. It is not a function of the transformation.

Why are these two seemingly unrelated concepts, weighted neighbor likelihood and conditional entropy, so closely related? We can gain some intuition about this equivalence by looking at the joint distribution of u and v. For any particular transformation we can sample points $[u(x_a), v(T(x_a))]^T$ and plot them. Figure 4.8 shows the joint samples of the signals from Figure 4.1 when aligned. The thin line in the plot is the weighted neighbor function approximation of this data; it is a good fit to the data. There is a noticeable clumping, or clustering, in the data. These clumps arise from the regions of almost constant intensity in the signals. There are four large regions of constant intensity and four clusters.

When these almost identical signals are aligned they are strongly correlated. Large values in one signal corresponds to large values of the other. Conversely, small values in one correspond to small values in the other. Correlation measures the tendency of the data to lie along the line $x = y$ (normalized correlation measures the tendency of the data to lie along some line of positive slope). Figure 4.9 shows the joint samples of the two signals shown in Figure 4.2. These signals are not linearly related or even correlated, but they are functionally related.

Weighted neighbor likelihood measures the quality of the weighted neighbor function approximation. In both of these graphs the points of the sample lie near the weighted neighbor function approximation. Moreover, in both of these graphs the joint distribution of samples is tightly packed together. Points are not distributed throughout the space, but lie instead in a small part of the joint space. This is the hallmark of a low entropy distribution.

We can generate similar graphs for signals that are not aligned. Figures 4.10 and 4.11 show the same signals except for the fact that the image has been shifted 30 units. For these shifted signals the structure of the joint distribution is destroyed. The weighted neighbor function

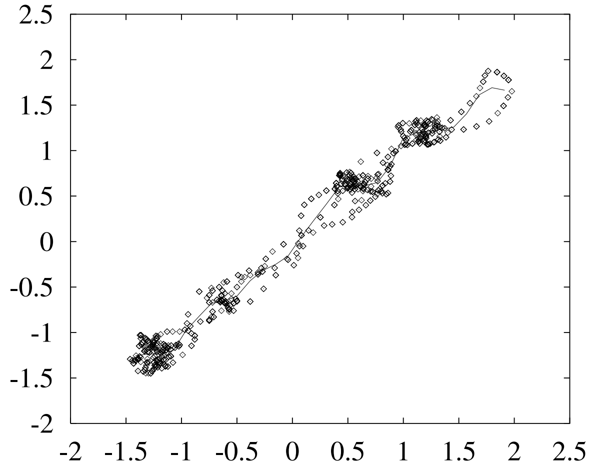


Figure 4.8: Samples from the joint space of $u(x)$ and $v(x) = u(x) + \eta$. A small black square is plotted for every pixel in the signals. The X-axis is the value of $u(x)$. The Y-axis is the value of $v(x)$. The clumping of points in clusters is caused by the regions of almost constant intensity in the images. The thin line plotted through the data is the weighted neighbor function estimate.

approximation is a terrible fit to the data. As a result the weighted neighbor likelihood of these signals is low.

Alternatively we could look directly at the distributions. When the signals are aligned the distributions are compact. When they are misaligned the distributions are spread out and haphazard. Or in other words, aligned signals have low joint entropy and misaligned signals have high joint entropy.

This suggests an alternative to weighted neighbor likelihood: the EMMA approximation of joint entropy. Graphed below are the EMMA estimates of joint entropy, $h^*(w)$, versus translation for each signal alignment problems discussed. Figure 4.12 shows a graph of joint entropy for the two signals that are nearly identical. Figure 4.13 shows a graph of joint entropy for the model and the non-linearly transformed image. In both case the graphs show strong minima at the correct aligning translation.

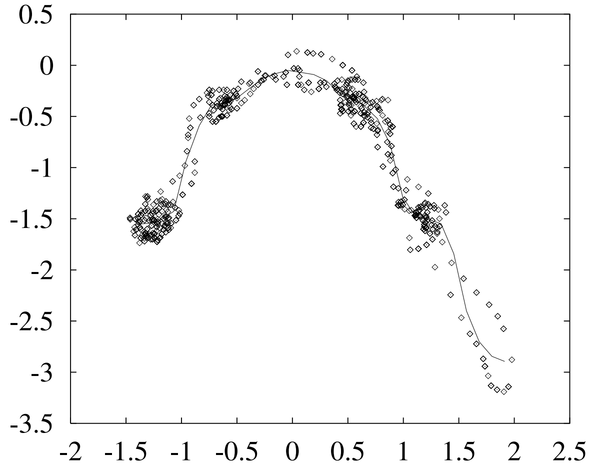


Figure 4.9: Samples from the joint space of $u(x)$ and $v(x) = -u(x)^2 + \eta$.

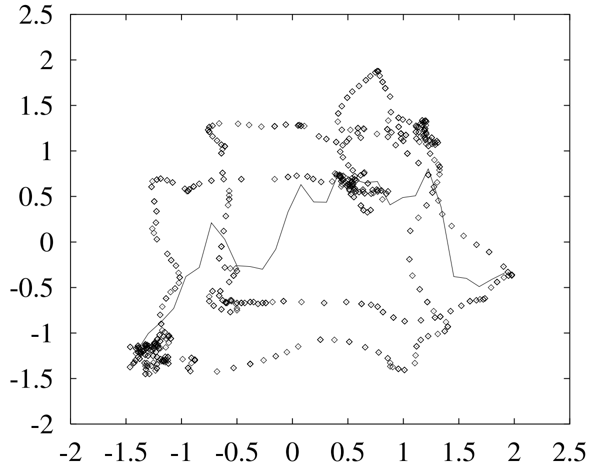


Figure 4.10: Samples from the joint space of $u(x)$ and $v(x) = u(x + 30) + \eta$. Unlike the previous graph these two signal are no longer aligned.

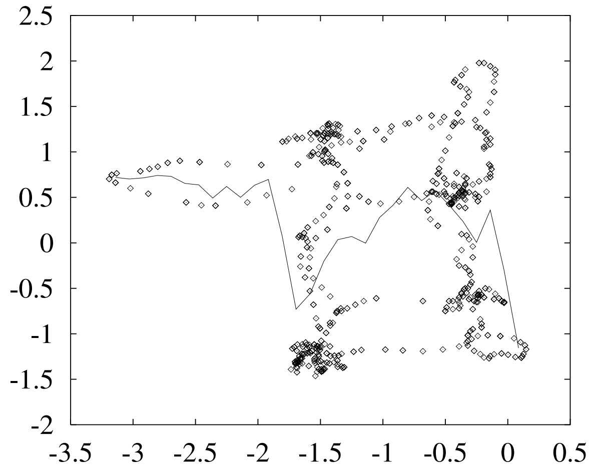


Figure 4.11: Samples from the joint space of $u(x)$ and $v(x) = -u(x+30)^2 + \eta$. The two signals are not aligned.

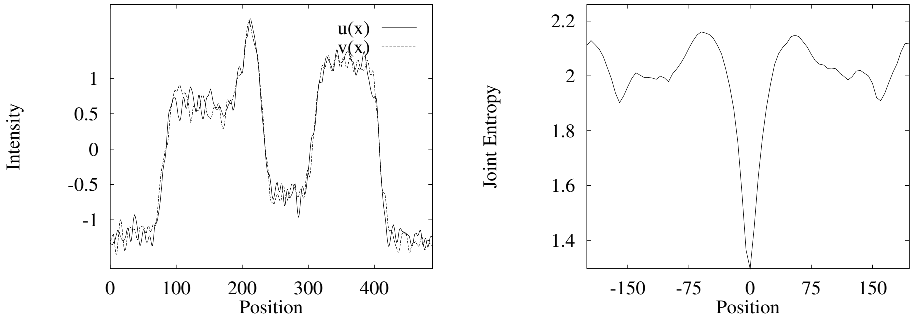


Figure 4.12: On the left is a plot of image and model that are identical except for noise. On the right is a plot of the estimated joint entropy versus translation.

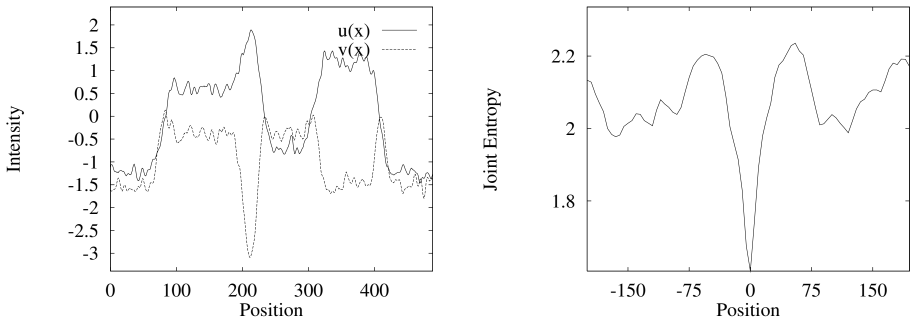


Figure 4.13: On the left is a plot of image and model that are related non-linearly. On the right is a plot of estimated joint entropy versus translation.

4.2 Weighted Neighbor Likelihood vs. EMMA

Weighted neighbor likelihood and EMMA are both smoothly differentiable functions that can be used to align signals when the imaging function is unknown. Qualitatively the EMMA estimate of joint entropy seems better. Joint entropy seems to have a wider basin in these synthetic experiments.

If weighted neighbor likelihood and EMMA are so similar, why is there a difference? Recall that weighted neighbor likelihood measures the conditional entropy of the image given the model. It does this under the assumption that the conditional distribution of the image is Gaussian. Weighted neighbor likelihood use the data around a point to estimate the mean of a Gaussian. The log likelihood of that point is then proportional to the squared difference from this mean. In general log likelihood calculations are very sensitive to *outliers*. Outliers are points that are, because of noise or measurement error, perturbed and land far from where they should have. Recall that the log likelihood of a sample is the sum of the log likelihoods of each point in the sample. As a result a single outlier can ruin a sample that would otherwise have had a high likelihood.

A more reasonable measure might introduce a bound on the penalty for a single point. Once a single point moved beyond a certain distance from the local mean the cost would no longer increase. Calculating likelihood in this way is closely related to the concept of a *robust* statistic

EMMA on the other hand does not assume that the conditional distribution of the image given the model is Gaussian. Instead it approximates the density non-parametrically. EMMA can handle situations where there are multiple peaks in the conditional distribution. While there is a likelihood penalty if a group of points are perturbed away from the local mean, it is not a function of the distance from the mean. Once these points move outside the effective range of the smoothing function there is no additional penalty. EMMA's robust nature prevents it from getting swamped by a few outliers in the joint distribution. This gives it a greater ability to deal with the distributions that arise from misalignments between model and image.

In the next sections we will describe a number of other situations where EMMA is better than weighted neighbor likelihood.

4.2.1 Non-functional Signals

Up until this point our analysis of alignment has assumed that there exists an imaging function that relates the model and the image. For at least two classes of problems there will be no imaging function at all. The first arises from a common situation in computer vision: occlusion. The second arises when the model does not contain all of the information required to predict the image. In both cases no single function, regardless of exogenous variables, can be used to predict the image from the model.

Figure 4.14 shows a graph of our original pair of signals, except that $v(x)$ has now been corrupted by an occlusion. Occlusion proves to be particularly bothersome for the alignment techniques we have proposed. For example, the basic assumption behind normalized cost has been violated; the occluded signal is not a linearly transformed version of the model. In addition, a quick glance at the joint space shows that the assumption behind weighted neighbor likelihood has also been violated (see Figure 4.15); even when the signals are aligned, there is no longer any function that relates $u(x)$ and $v(x)$. Figure 4.16 show a graph of weighted neighbor likelihood versus translation. The global minimum no longer coincides with the correct translation.

In some cases EMMA can be used to align partially occluded signals. Joint entropy does not suffer from the strong assumption that the signals are functionally related. Though part of the signal may be corrupted, the remaining parts retain their low entropy relationship. Figure 4.17 show a plot of joint entropy for the occluded pair of signals.

The simplest example of non-functional signals often arises when the model and the image are swapped. Whenever the function between the model and the image is non-monotonic, the relationship between the image and the model is non-functional. The non-monotonically related signals shown in Figure 4.2 are an example. Figure 4.18 shows the joint space of the swapped signals and a weighted neighbor function approximation. The function fit to this joint space is a terrible approximation of the data. The quality of the function approximation points out an important limitation of weighted neighbor likelihood. While normalized cost is a *symmetric* comparison metric, weighted neighbor likelihood is *not*. It may seem at first that this is an unimportant distinction. It is not. Symmetric measures allow us to match images to models as well as models to images. This can be critical when it is not possible to construct a detailed model.

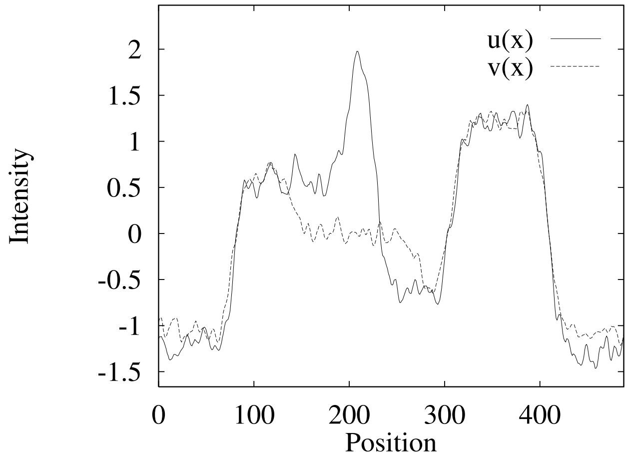


Figure 4.14: Graph of $u(x)$ and $v(x)$ where $v(x)$ has been perturbed by noise, and a portion of it occluded.

Both joint entropy and mutual information are symmetric measures. A plot of EMMA entropy for the swapped signals is identical to Figure 4.13.

A more complex example of non-functional signals arises when both the model and the image are functions of some third unmeasurable signal. Call this signal $z(x)$. There are now two imaging functions, one that creates the model $u(x) = F_u(z(x), q_u)$, and another that creates the image $v(T(x)) = F_v(z(x), q_v)$. Medical registration, which we describe in some detail in the next chapter, is a clear example of a two sensor problem. In medical registration one seeks an alignment of signals from two types of sensors (for example a CT scan and an MRI scan). Neither gives perfect information about the object, and neither is completely predictable from the other.

The two sensor problem can be simulated by transforming our original signal by two different non-linear transforms. Using the original signal from Figure 4.1 we can define $F_u(z) = sin(2z)$ and $F_v(z) = z^2$. The resulting aligned distribution should fall approximately along a line that looks like Figure 4.19. The actual distribution is shown in Figure 4.20. Here too EMMA shows a strong minimum at the correct alignment of model and image (see Figure 4.21).

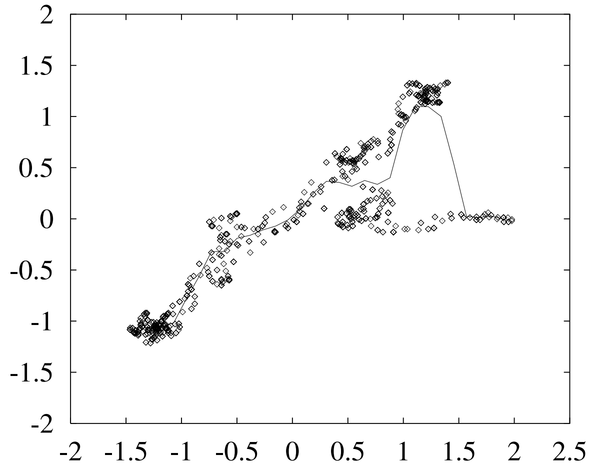


Figure 4.15: Samples from the joint space of $u(x)$ and $v(x)$ where the image has been occluded. Though these signals are aligned the the weighted neighbor function is a terrible fit to the data. The data is segregated into two parts: the linearly related part that arise from the non-occluded signal, and the constant part that projects to the occluded part.

Mutual Information versus Joint Entropy

Recall that conditional entropy is not a measure of the dependence between two signals (see Section 2.2). Conditional entropy $h(v|u)$ can be small for two different reasons: $h(v)$ is itself small, or v is dependent on u. Mutual information is a better measure of dependence. For the simple examples described in this chapter $h(v \mid u)$ can be used alone as a measure of alignment. In more complex examples, where the model can change scale or project on a limited part of the image, $h(v)$ must be taken into account. In general we will solve alignment problems by maximizing mutual information.

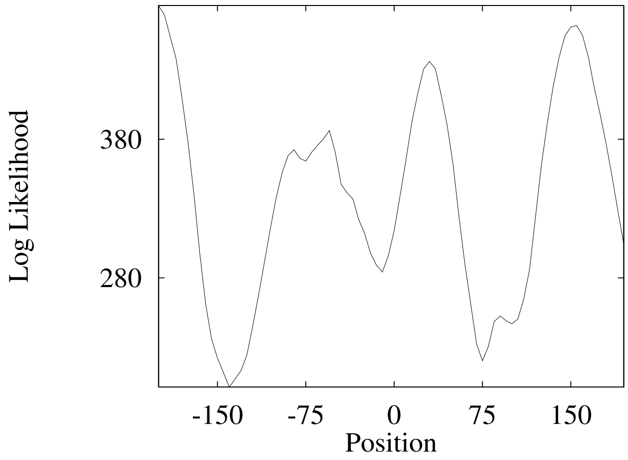


Figure 4.16: Graph of weighted neighbor likelihood versus translation where $v(x)$ has been occluded.

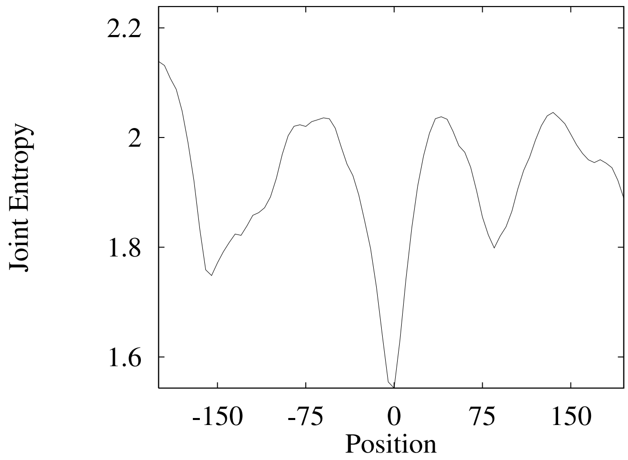


Figure 4.17: Graph of EMMA joint entropy estimate versus translation for the occluded pair of signals.

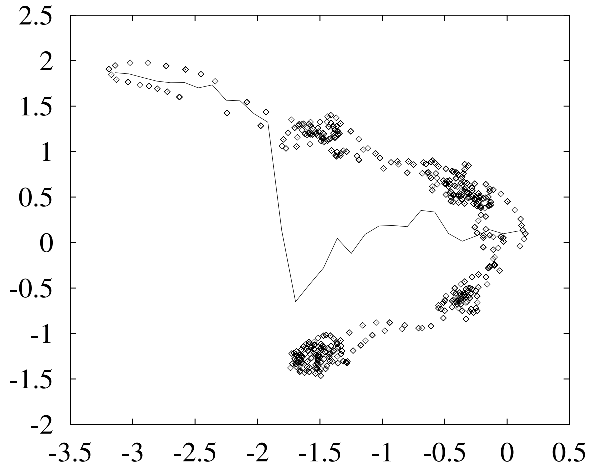


Figure 4.18: Samples from the joint space of $u(x)$ and $v(x) = u(x)^2 + \eta$. The roles of the model and the image have been reversed.

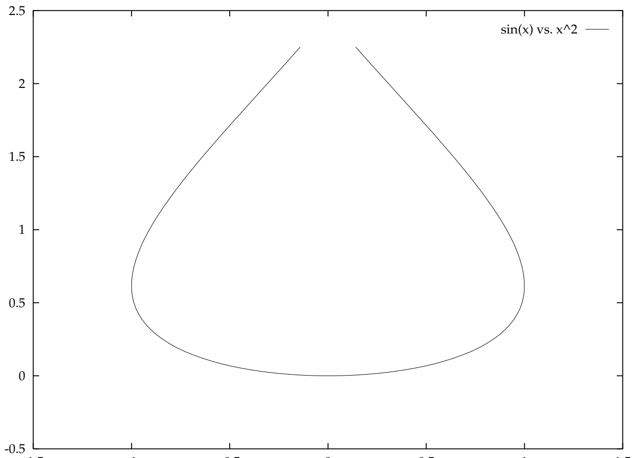


Figure 4.19: Graph of the function defined by $u = sin(2z)$ and $v = z^2$ as z varies from -1.5 to 1.5.

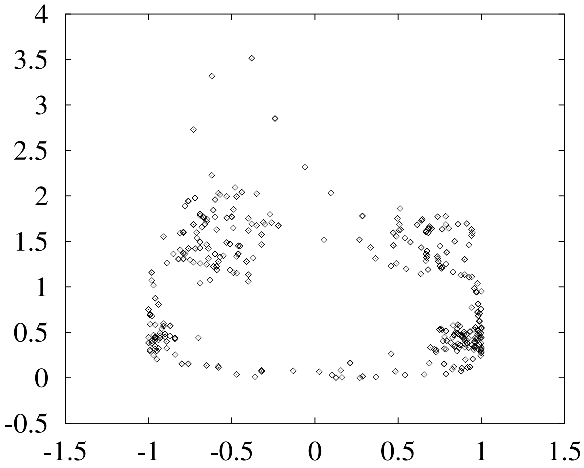


Figure 4.20: Samples from the joint space of from the simulated two sensory data.

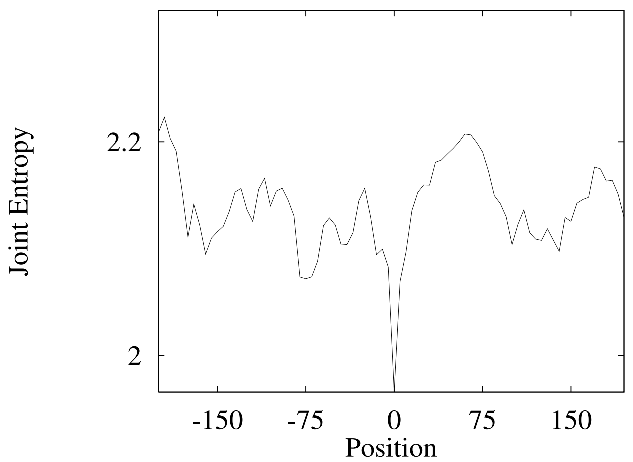


Figure 4.21: Graph of EMMA joint entropy estimate versus translation for the non-functional pair of signals.

4.3 Alignment Derivation

Let us derive the equations and algorithms used for alignment by maximization of mutual information. We wish to maximize the mutual information between model and image signals. This requires a search of the aligning transformation space. We will use the stochastic gradient descent algorithm described in Section 3.3.

The derivative of mutual information is

$$\frac{d}{dT}I(u(x), v(T(x))) = \frac{d}{dT}h(u(x)) + \frac{d}{dT}h(v(T(x))) - \frac{d}{dT}h(u(x), v(T(x))) \ .$$

Since $h(u(x))$ is not a function of T, it drops from our calculations. Using the EMMA entropy estimate,

$$\frac{d}{dT}I(u(x), v(T(x))) \approx \frac{1}{N_B}\sum_{x_i \in B}\sum_{x_j \in A} W_v(v_i, v_j)\,(v_i - v_j)^T \psi_v^{-1}\frac{d}{dT}(v_i - v_j) \tag{4.33}$$

$$-\frac{1}{N_B}\sum_{x_i \in B}\sum_{x_j \in A} W_{uv}(w_i, w_j)\,(w_i - w_j)^T \psi_{uv}^{-1}\frac{d}{dT}(w_i - w_j) \ . \tag{4.34}$$

The following definitions have been used:

$$W_v(v_i, v_j) \equiv \frac{g_{\psi_v}(v_i - v_j)}{\sum_{x_k \in A} g_{\psi_v}(v_i - v_k)} \ ,$$

$$W_{uv}(w_i, w_j) \equiv \frac{g_{\psi_{uv}}(w_i - w_j)}{\sum_{x_k \in A} g_{\psi_{uv}}(w_i - w_k)} \ ,$$

$$u_i \equiv u(x_i)\ ,\ u_j \equiv u(x_j)\ ,\ u_k \equiv u(x_k)\ ,$$

$$v_i \equiv v(T(x_i))\ ,\ v_j \equiv v(T(x_j))\ ,\ v_k \equiv v(T(x_k))\ ,$$

$$w_i \equiv [u_i, v_i]^T\ ,\ w_j \equiv [u_j, v_j]^T\ ,\ \text{and}\ w_k \equiv [u_k, v_k]^T \quad .$$

We assume that the covariance matrices of the component densities used in the approximation scheme for the joint density are block diagonal,

$$\psi_{uv}^{-1} = \mathrm{DIAG}(\psi_{uu}^{-1}, \psi_{vv}^{-1}) \ ,$$

and we obtain an estimate for the derivative of the mutual information as follows

$$\widehat{\frac{dI}{dT}} = \frac{1}{N_B} \sum_{x_i \in B} \sum_{x_j \in A} (v_i, v_j)^T [W_v(v_i, v_j)\psi_v^{-1} - W_{uv}(w_i, w_j)\psi_{vv}^{-1}] \frac{d}{dT}(v_i - v_j) \ .$$

If we are to increase the mutual information, then the effect of the first term in the brackets may be interpreted as acting to increase the squared distance between pairs of samples that are nearby in image intensity, while the second term acts to decrease the squared distance between pairs of samples that are nearby in *both* image intensity and the model properties. It is important to emphasize that distances are in the space of values (intensities, brightness, or surface properties), rather than coordinate locations.

The term $\frac{d}{dT}(v_i - v_j)$ will generally involve gradients of the image intensities and the derivative of transformed coordinates with respect to the transformation. In the simple case that T is a linear operator, we obtain the following outer product expression,

$$\frac{d}{dT} v(T(x_i)) = \nabla v(T(x_i)) x_i^T \ .$$

4.4 Matching and Minimum Description Length

There is another entirely different motivation for using mutual information as a alignment metric. Alignment, and many other vision problems, can be reformulated as *minimum description length* (MDL) problems (Rissanen, 1978; Leclerc, 1988). MDL can provide us with some new insight into the problem of alignment and help us derive a missing and often useful term in the alignment equations.

The standard framework of MDL involves a sender and a receiver communicating descriptions of images. Given that the sender and the receiver have an agreed upon language for describing images, the sender's goal is to find a message that will accurately describe an image in the fewest bits. The concept of description length is clearly related to the code length introduced in Section 2.2.

For the problem of alignment we will assume that the sender and the receiver share the same set of object models. The sender's goal is to communicate an image of one of these models. Knowing nothing else, the sender could simply ignore the models and send a message describing the entire image. This would require a message that is on average as long as the entropy of the image. However, whenever the image is an observation of a model a more efficient approach is possible. For example the sender could send the pose of the model and a description for how to render it. From these the receiver can reconstruct the part of the original image in which the model lies. To send the entire image, the sender need only encode the errors in this reconstruction, if there are any, and any part of the image that is unexplained by the model. Alignment can be thought of as the process by which the sender attempts to find the model pose that minimizes the code length of the overall message.

The encoding of the entire image has several parts: (1) a message describing the pose; (2) a message describing the imaging function; (3) a message describing the errors in the reconstruction; and (4) a message describing the parts of the image unexplained by the model. The length of each part of the message is proportional to its entropy. We can assume that poses are uniformly distributed, and that sending a pose incurs some small uniform cost. The length of part (4) is the entropy of the image that is unexplained. Parts (2) and (3) can be interpreted in two ways. We can assume that the imaging function can be sent with a fixed or small cost. Part (3) is then proportional to the conditional entropy of the image given the model and imaging function. This is precisely what was estimated and minimized with weighted neighbor alignment. A second interpretation comes from EMMA. EMMA estimates the joint entropy of the model and image, $h(u, v)$. The conditional entropy of the image given the model can be computed as $h(v|u) = h(u, v) - h(u)$. Since the entropy of the model is fixed, minimizing the joint entropy minimizes the conditional entropy. In both cases entropy based alignment as proposed in the first part of this chapter minimizes the cost of sending parts (1), (2) and (3). MDL suggests that we must also minimize the entropy of the unmodeled part of the image.

In the previous information theoretic formulation there was no concept of pixels or of the proportion of the image explained by the model. In fact, in the previous formulation the entropy of the explained part of the image could get larger as the model shrunk. For example, assume that the model covers a contiguous region of an image where most of the pixels have constant value. At the center of this region is a small patch containing varied pixels. Recall that the image is sampled at points that are projected from the model. Most of the model points will project into the region of constant intensity and a few will project onto the varied patch. The resulting distribution of image pixels, because it has many samples of the same value, has fairly low entropy. If the model were shrunk to cover only the varied patch, then all of the points from the model would fall in the varied region. The new distribution of pixel values will have higher entropy.

Can the mutual information between the model and the entire image be measured? This would be a more direct parallel to the MDL framework. MDL directs the sender to transmit the model and pose that will produce an encoding for the entire image that is shortest. As the model explains more, less of the image needs to be encoded directly. We can derive an approximation to mutual information between the model and the entire image. Let U and V represent the whole model and whole image respectively. The mutual information equation can now be rewritten,

$$I(U,V) = h(U) + h(V) - h(U,V) \ . \tag{4.35}$$

The dependence of $h(U,V)$ on T is implicit in the above equation.

Logically, the image can be split into two parts: $V_{\bar{m}}$ and V_m. V_m is the part of the image in which the model lies. $V_{\bar{m}}$ is the unmodeled part of the image. We can define them as

$$V_m \equiv \{v(y) \text{ such that } u(T^{-1}(y)) \text{ is defined}\} \ , \tag{4.36}$$

and,

$$V_{\bar{m}} \equiv V - V_m \ . \tag{4.37}$$

It is assumed that the two parts of the image are independent. This allows us to split the entropies that involve V into two distinct parts,

$$\begin{aligned}
I(U,V) &= h(U) + h(V) - h(U,V) && (4.38)\\
&= h(U) + h(V_m, V_{\bar{m}}) - h(U, V_m, V_{\bar{m}}) && (4.39)\\
&= h(U) + h(V_m) + h(V_{\bar{m}}) - h(U, V_m) - h(V_{\bar{m}}) && (4.40)\\
&= h(U) + h(V_m) - h(U, V_m) \ . && (4.41)
\end{aligned}$$

Step (4.40) relies on the assumption that $V_{\bar{m}}$ is independent of both V_m and U (i.e. that the background is independent both of the image of the object and the object). Equation 4.41 directs us to maximize the entropy of the modeled part of the image, $h(V_m)$, as well as minimizing the joint entropy of model and the image.

EMMA can estimate the entropy of a signal, but what is the entropy of an entire image? This is a very difficult question to answer. The entropy of an image is a function of the distribution of a vector random variable with between ten thousand and a million dimensions. Though the entropy of an image can be modeled as the the sum of the pixel entropies, this is guaranteed to be an overestimate of the true image entropy.

One of the problems with mutual information alignment is that it does not tell us whether the object is present in the image or not. The principle of minimum description length should allow us to derive a decision threshold for the presence of an object. The object is present if the image is easier to encode with it than without it. Unfortunately this decision is highly dependent on the estimate of the entropy, or code length, of the explained part of the image. The naive overestimate of image entropy—that simply sums the entropies of the pixels—is not tight enough to determine the decision threshold correctly. An important area of open research is deriving a more reasonable estimate of the code length of an image.

In previous derivations points were sampled from the model and projected into the image. Since we are now explicitly modeling the entropy of the image, we will sample from the image and project back into the model. The joint entropy is then,

$$h(v(y), u(T^{-1}(y))) = E_Y[log\ p(v(y), u(T^{-1}(y)))] \tag{4.42}$$
$$= \int p(v(y), u(T^{-1}(y)))log\ p(v(y), u(T^{-1}(y)))dy \tag{4.43}$$
$$= \int p(v(T(x)), u(x))log\ p(v(T(x)), u(x))A(T, x)dx \tag{4.44}$$
$$= E_X[A(T, x)log\ p(v(T(x)), u(x))]\ . \tag{4.45}$$

Equation (4.44) involves a change of variables from y back into x. The correcting term $A(T, x)$ plays the role of the Jacobian of the transformation, measuring the ratio of the area in the model as it projects into the image. For affine transformations the projected area is related to the model area by the determinant of the transformation. For projective transformations there will also be a term that arises from foreshortening. The mutual information is then

$$I(U, V) \approx E_b\left[A(T, x_b)\begin{pmatrix} +log(P^*(u(x_b), a)) \\ +log(P^*(v(T(x_b)), a)) \\ -log(P^*(\{v(T(x_b)), u(x_b)\}, a)) \end{pmatrix}\right]\ . \tag{4.46}$$

The mutual information between the whole model and image is the EMMA estimate of the mutual information between the model and image signals weighted by the projected area of the model. This new formulation of mutual information includes a term which encourages the model to explain as much of the image as possible. The derivative

$$\frac{d}{dT} I(U,V) \approx E_b \left[\begin{array}{l} A(T,x_b)\frac{d}{dT}\left(\begin{array}{l} +log(P^*(u(x_b),a)) \\ +log(P^*(v(T(x_b)),a)) \\ -log(P^*(\{v(T(x_b)),u(x_b)\},a)) \end{array} \right) \\ + \left(\begin{array}{l} +log(P^*(u(x_b),a)) \\ +log(P^*(v(T(x_b)),a)) \\ -log(P^*(\{v(T(x_b)),u(x_b)\},a)) \end{array} \right) \frac{d}{dT}A(T,x_b) \end{array} \right] , \qquad (4.47)$$

encourages the model to grow as large as possible.

4.5 Summary

In this chapter several motivations for the selection of mutual information as a measure of alignment have been presented. The primary existing measure of alignment, correlation, is rederived as a maximum likelihood technique which has a number of important weaknesses. The concept of maximum likelihood matching is then extended to define a more general measure of alignment. This measure, called weighted neighbor likelihood, can be interpreted as estimating the conditional entropy of an image given a model. While it is more general than correlation, weighted neighbor likelihood is still limited. It can be still further generalized yielding a technique called EMMA alignment. This technique explicitly estimates the mutual information between an image and model. A number of synthetic experiments demonstrate that mutual information is a very flexible measure of alignment.

In the final section of this chapter the concept of minimum description length is used as yet another motivation for mutual information as an alignment measure. This alternative framework encourages the model to explain as much of the image as possible.

Chapter 5

Alignment Experiments

This chapter contains a number of experiments designed to demonstrate that alignment by maximization of mutual information is a practical technique. The previous chapter contained a very general definition of alignment. Though a procedure for adjusting the pose parameters during alignment was derived, many of the details regarding representations and implementations were left out. Each experiment described in this chapter will include both these needed details and a general discussion of the experimental framework.

By the end of this chapter we will have developed some familiarity with the application of EMMA alignment. The chapter will conclude with a section describing explicit limitations of this approach. In addition to describing problems for which EMMA alignment is poorly suited, it will be emphasized that EMMA alignment by itself is not a complete object recognition system.

For clarity the parameters and assumptions that underly EMMA alignment have been identified. We have broken the process of setting up an experiment into discrete steps that can be applied to a wide variety of alignment problems (see Table 5.1). Along with the description of each experiment we will include a similar table with a specific realization for each step.

1.	Choose a model and image representation (i. e. define $u()$ and $v()$). Define an interpolation scheme for sampling $v()$ at non-integral coordinates.
2.	Choose a scheme for sampling the model (i.e. define x).
3.	Determine the space of possible aligning transformations and its concrete representation (i.e. define T). The definition of the random variables $u(x)$ and $v(T(x))$ is now complete.
4.	Derive an expression for $dv(y)/dy$.
5.	Pick a metric for computing distances between pairs of samples of $u(x)$, $v(T(x))$, and $\{u(x), v(T(x))\}$.
6.	Pick the variance for the component densities: ψ.
7.	Choose a value for p_{min}.
8.	Determine the number of samples used to estimate the distribution, and the number used to estimate the entropy.
9.	Pick a parameter update rate, λ. In general the update rate will decrease with time.

Table 5.1: The process for setting up an alignment.

5.1 Alignment of 3D Objects to Video

In our first experiment we will return to the example described in the introduction: alignment of a three-dimensional object to a video image. In all of our alignment experiments we will assume that the entire object has the same surface properties. We can then treat surface property as yet another exogenous variable.

Following Table 5.1:

1. Models are a collection of points that lie on the surface of the object. We chose this representation because it is capable of representing any shape including smoothly curved or irregular forms. It is equally capable of representing objects with flat faces such as polyhedra. The models have been constructed so the distribution of surface points is as close to uniform as possible. Associated with each surface point is the local surface normal, a unit vector perpendicular to the surface. The models used have between 7000 and 65,000 points. Video images are represented as simple two dimensional arrays.

2. The random variable x that is used to sample the model and image is defined from the

model. A trial of x is a randomly selected model point. The value of the trial is the 3D location of that model point. We sample the points of the model uniformly.

3. The transformation space is the space of rigid three dimensional translations and rotations followed by perspective projection. The overall transformation is a concatenation of rotations and translations each acting on the pre-defined "center" of the object.

   Because of self-occlusion, not every point on the model is visible. Visibility is determined by a Z-buffer rendering of the model. Z-buffer rendering takes each point in the model and projects it into the image. When multiple points fall onto the same pixel, only the point that is nearest is considered visible. As pose changes, some points become visible and others become invisible. In theory Z-buffering needs be repeated every time the pose of an object changes. Unfortunately, Z-buffering takes time proportional to the size of the model. This cost is far larger than the cost of computing an estimate for the derivative of entropy. Since pose does not change much between iterations of gradient descent, it has proven sufficient to Z-buffer every 300 iterations.

4. The derivative $dv(y)/dy$ is the spatial gradient of the image.

5. The metric used for comparing points sampled from the image is squared difference. The representation of joint events, $w = \{v(T\ x), u(x)\}$, is somewhat more complex. We will represent only two dimensions of the normal vector: the x and y components. Since the normal is always a unit vector, the z component is redundant. The joint events are therefore three dimensional vectors, two components from the model and one from the image. We will use Euclidean distance to measure the distance between joint events.

6. Since we will be using diagonal covariance matrices for the smoothing functions, four variances are required. Three for the joint entropy and one for the image entropy. Based on maximum likelihood estimates from aligned objects, we have settled on a single set of smoothing parameters that we will use for all of our 3D alignment experiments. For the joint entropy, the variance of x and y components of the normal are both 0.3 and the variance for image intensity is 0.2. For the image entropy, the variance for image intensity is 0.15. Having a single set of parameters is for every experiment is possible in part because we have pre-normalized all images so that their variance is 1.0.

7. We will use a value of 0.01 for p_{min}. The alignment process shows very little sensitivity to p_{min}. We have repeated a number of experiments with a p_{min} value of 0.1 and

1.0. Our results are not significantly different. Values that are more than a factor of 10 smaller than 0.01 cause the derivative of estimated entropy to be too noisy (see Section 3.3). This noise can prevent convergence to the correct pose.

8. Rather than draw two different samples, we will use the cross-validation approximation (see Section 2.4.3). In all of our experiments we use a sample size of 25.

9. Finally, we must choose a parameter update rate, λ. Actually, since the units of rotation and translation are very different two update rates are necessary. Internally we represent rotations in radians and translations in millimeters. For an object with a 100 millimeter radius a rotation of 0.01 radians about the center of mass can translate a model point up to 1 millimeter. A translation of 0.01 can at most translate a model point 0.01 millimeters. The derivative of mutual information with respect to a model point's position is a combination of a rotation and translation. A small step in the direction of the derivative will move the model point up to 100 times further by rotation than translation. If there is only a single update rate a poor compromise must be made between the rapid changes that arise from the rotation and the slow changes that arise from translation. If the rotation update rate is reduced by a factor of 100 the model point will move approximately as far by rotation as it does by translation. Scale issues such as these do not arise when more complex gradient descent techniques are used, for example conjugate gradient descent or Levenberg-Marquardt. Unfortunately, neither of these techniques can use stochastic estimates of the gradient. Since our models have a radius that is on the order of 100 millimeters, we have chosen rotation update rates are 100 times smaller than translation rates. Most of our 3D alignment experiments proceed in two stages. In the first stage the rotation update rate is 0.0005 and the translation update rate is 0.05. After a number of iterations the update rates are then reduced to 0.0001 and 0.01 respectively. We have chosen a simple automatic descent procedure in an effort to simplify subsequent analysis of convergence.

The realization of the basic framework is summarized in Table 5.2.

1.	Define the model and image $u()$ and $v()$: $u()$ contains points distributed on the surface of the object. Each point has an associated normal. $v()$ is the image of intensities.
2.	Sampling x: The sampling is determined by the distribution of surface points which is close to uniform.
3.	Transformation space T: The space of rigid 3D rotations and translations by perspective projection using an estimate for the camera parameters.
4.	Definition of $dv(y)/dy$: This is the intensity gradient.
5.	Distance metric: Euclidean distance.
6.	Variance, ψ: Assuming diagonal covariance matrices, four different variance are necessary, three for the joint entropy estimate and one for the image entropy estimate. The variances were 0.3, 0.3, and 0.2 for the x component of the normal, y component of the normal, and image intensity. The variance was 0.15 for the image entropy.
7.	Minimum probability, p_{min}: 0.01.
8.	Number of samples: One sample of 25 using cross-validation.
9.	Update rate, λ: Rotation rate: 0.0005 for 3200 steps and then 0.0001 for 3200 steps. Translation rate: 0.05 and then 0.01.

Table 5.2: Summary of 3D to video alignment.

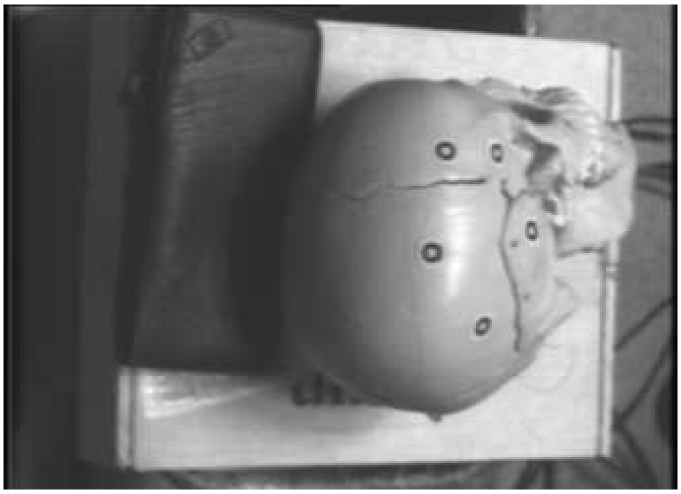


Figure 5.1: A typical image of the skull object.

5.1.1 Alignment of Skull Model

In our first experiment we will align a 3D skull model to a number of different images. The skull model was produced automatically from a Computed Tomography (CT) scan of a plastic skull[1]. The same plastic skull was then photographed in a number of different poses under natural lighting[2]. The skull model contains 65000 points. The video images are 240 by 320 pixels. Figure 5.1 is an example video image of the skull.

Figure 5.2 contains a representation of the shape of the skull model. It is an image displaying the distance from the camera to the visible points on the skull model. White is further and black nearer. This image is computed by projecting each model point into the image plane. The pixel to which the model point projects records the distance of the model point from the camera. There may, however, be a number of model points that project to the same image pixel. In this case, the depth of the model point which is nearest the camera is used. Since the model is constructed from a collection of points, it is not dense. As a result there are some pixels to which no model point projects. A few of these pixels, which remain white, appear throughout the model.

[1]Thanks to Gil Ettinger and Ron Kikinis for providing the skull model. Their work on medical registration using this model is described in (Grimson et al., 1994).

[2]Thanks to J. P. Mellor for providing the skull images. His work on registration is described in (Mellor, 1995).

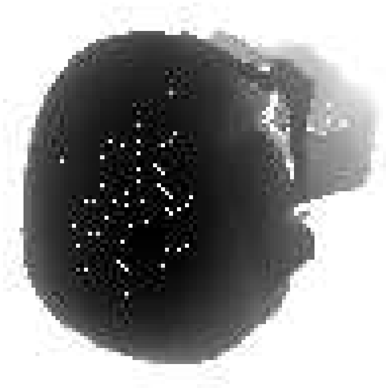


Figure 5.2: A depth map of the skull model. See text for description.

We can use a Lambertian reflectance model to render a graphical picture of the skull. A Lambertian model relates the model normal and image intensity:

$$v(Tx) = \sum_i \alpha_i \vec{l_i} \cdot u(x) \quad , \tag{5.1}$$

where the model value $u(x)$ is the normal vector of a surface patch, l_i is a vector pointing toward light source i, and α_i is proportional to the intensity of that light source. Figure 5.3 shows a rendered version of the model in the same pose as Figure 5.1. To the human eye this sort of image is more readily interpretable than a depth map. We can bring to bear our substantial visual competence when the shape of an object is rendered as an image. From Figure 5.3 it is almost immediately clear that the pose of the object model is close to correct. There is however no simple relationship between the intensities of the video image and the rendered image.

The goals of this first experiment is to answer three questions. (1) Can EMMA align a complex 3D object model to a number of different images taken under uncontrolled lighting? (2) How long does EMMA alignment take to run? (3) What is the range of poses from which a "correct" alignment can be obtained. Regarding this third point, we do not have true information about either the pose of the object nor the camera parameters of the video camera. The "correct" pose has been determined by inspection of the alignment results. We can however ask a related question about reliability. How far can the the object be perturbed

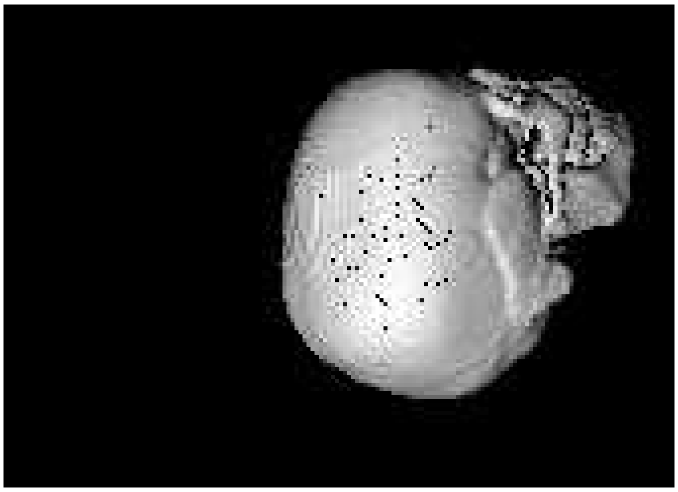


Figure 5.3: A rendered image of the skull model.

away from the "correct" pose and have EMMA alignment reliably re-align it?

To answer our first question, we must establish that in the six dimensional space of rigid transformations there is a maximum of mutual information at a plausible alignment pose. For each image the object model was initially adjusted so that it's pose was close to correct. This was done by eye. EMMA alignment was then used to pull the object into a "correct" pose.

One scheme for assessing the quality of an alignment is to display the model pose and the video image together. This can be done by taking a random collection of model points and projecting them into the coordinate frame of the image. The pixels to which the model points project are then set to white. The nature of the alignment is readily apparent from such images. When the model and image are misaligned model points will project onto the background and the coverage of the object's image will be incomplete. When the model and image are correctly aligned there is close agreement between the occluding contours of the model points and the object's image. Figure 5.4 shows an initial incorrect pose in this way. Figure 5.5 shows the final pose obtained after running EMMA alignment.

Figures 5.6, 5.7, and 5.8 show the final alignment obtained for three other images. Notice that in each of these images the boundaries of the skull are in close agreement with the outline of the model points.

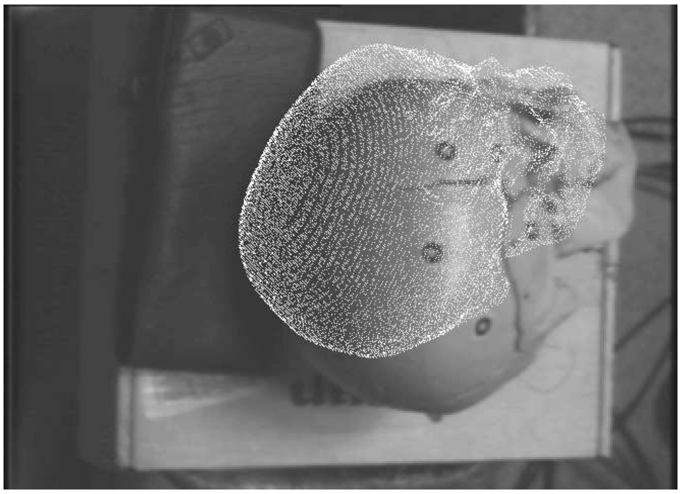


Figure 5.4: Initial pose of the skull model before alignment.

We would like to emphasize that in none of these experiments have we pre-segmented the image. The initial poses often project the model into regions of the image that contain a significant amount of clutter. EMMA reliably settles on a pose where few if any of the model points project onto the background.

In answer to the second question, EMMA requires roughly 35 seconds on a Sun SparcStation5 for each of the alignments shown above. Run times are identical because we have chosen to use a fixed number of update iterations for each alignment experiment. In some cases an accurate alignment was obtained well before the full number of iterations had been completed. In others it appeared that the final alignment could have been improved if the number of iterations were increased.

There are few if any principled results on the convergence of stochastic approximation. Convergence detection is a subtle issue. For example, EMMA does not make a direct estimate of the mutual information between model and image. During alignment only a stochastic estimate of the gradient is available. It may be possible to construct an ad hoc procedure that would be able to detect convergence. Alignment could then be continued until the pose estimate had converged.

From an analysis of the program's memory access and computation patterns, we conclude that an implementation on a digital signal processor would be as much as 100 times faster than our current implementation. One major issue is cache performance. Because EMMA

Figure 5.5: Final pose of the skull model after alignment.

randomly accesses each of the points in the image and model, much time is wasted flushing and refilling the cache. The cache on a general purpose processor is often fairly limited. Most digital signal processors include a large quantity of fast SRAM, eliminating the need for a cache. For random memory accesses a digital signal processor should be approximately 5 times faster than a conventional computer. The inner loop of the EMMA derivative estimation procedure is dominated by simple floating point operations. Modern digital signal processors can execute these instructions 10 to 20 times faster than conventional computers. Together these advantages should lead to an overall improvement in speed of between 50 and 100.

A number of randomized experiments were performed to determine the reliability, accuracy and repeatability of alignment. This data is reported in Table 5.3. An initial alignment was performed to establish a base pose. This pose, shown in Figure 5.5, is used as a point of reference. A set of randomized experiments was performed where the base pose is first perturbed, and then EMMA is used to re-align the image and model. The perturbation is computed as follows: a random uniformly distributed offset is added to each translational axis (labeled ΔT) and then the model is rotated about a randomly selected axis by a random uniformly selected angle ($\Delta\theta$). There were four experiments each including 50 random initial poses. The distribution of the final and initial poses can be compared by comparing the variance of the location of the centroid, computed separately in X, Y and Z. Furthermore, the average angular rotation from the true pose is computed (labeled $\overline{|\triangle\theta|}$). Finally, the number of poses that failed to converge near the correct solution is reported. The final statistics are

Figure 5.6: Final pose of the skull model after alignment.

$\triangle$T	$\triangle\theta$	INITIAL				FINAL				
X Y Z		σ_X	σ_Y	σ_Z	$\overline{\mid\triangle\theta\mid}$	σ_X	σ_Y	σ_Z	$\overline{\mid\triangle\theta\mid}$	
±mm	°	mm			°	mm			°	%
10	10	5.94	5.56	6.11	5.11	.61	.53	5.49	3.22	100
30	10	16.53	18.00	16.82	5.88	1.80	.81	14.56	2.77	96
20	20	10.12	12.04	10.77	11.56	1.11	.41	9.18	3.31	96
$[10, 20]$	$[20, 40]$	14.83	15.46	14.466	28.70	1.87	2.22	14.19	3.05	78

Table 5.3: Skull Results Table. The final column contains the percentage of poses that successfully converged to a pose near the correct pose.

only evaluated over the poses that converged near the correct solution.

These experiments demonstrate that the alignment procedure is reliable when the initial pose is close to the "correct" pose. Outside of this range gradient descent, by itself, is not capable of converging to the correct solution. The capture range is not unreasonably small however. Translations as large as half the diameter of the skull can be accommodated, as can rotations in the plane of up to 45 degrees. Empirically it seems that alignment is most sensitive to rotation in depth. This is not terribly surprising since only the visible points play a role in the calculation of the derivative. As a result, when the chin is hidden the derivative gives you no information about how move the chin out from behind the rest of the skull.

Finally, we have done a number of experiments to demonstrate that EMMA alignment

Figure 5.7: Final pose of the skull model after alignment.

can deal with occlusion. Figure 5.9 shows an initial and final alignment for an image that includes an artificial occlusion that covers the entire chin area. The final alignment is very close to the correct one despite the occlusion. Figure 5.10 shows an initial and final pose for a more complex occlusion. In this image we have replaced a rectangular window with another randomly chosen window of the image. The source of the rectangle is near the bottom of the image. In a number of experiments, we have found that alignment to occluded images can require more time for convergence.

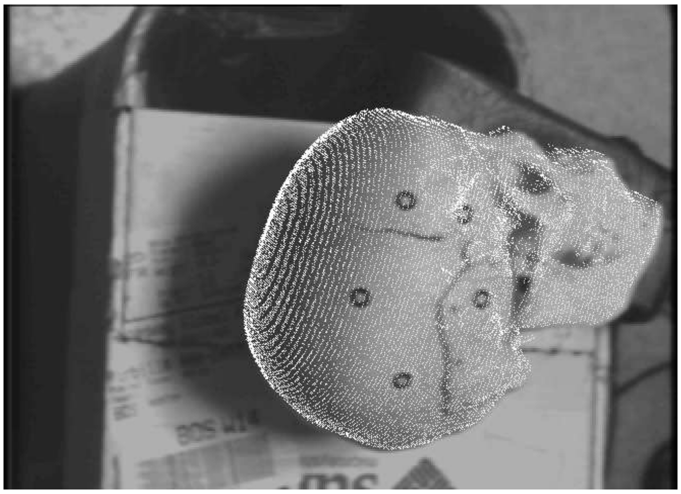


Figure 5.8: Final pose of the skull model after alignment.

Figure 5.9: An image including an artificial occlusion. White spots denote the pose of the model. On the left is the initial pose, on the right is the final pose.

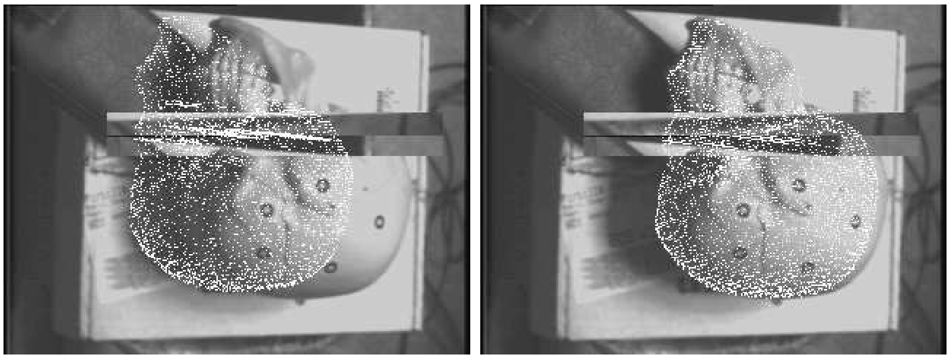


Figure 5.10: An image including an artificial occlusion. White spots denote the pose of the model. On the left is the initial pose, on the right is the final pose.

5.1.2 Alignment of Head Model

We have repeated many of the skull experiments with a three dimensional model of a human head. This model was obtained from a Cyberware scan of the subject that was taken approximately two years before the video images[3]. A Cyberware scan is a complete three dimensional representation of the shape of the subject's head in cylindrical coordinates. The surface normals were computed from the surface by smoothing and differencing neighboring surface points.

The experiments in this section are designed to answer two questions: (1) Will the same techniques and parameters work with two different types of models and images? (2) Is it possible to use the pose refinement procedure to track a moving object in a video sequence? Figure 5.11 shows an image of the head and a rendering of the model.

How are the face experiments different from the skull experiments? Firstly, the face model is much smoother than the skull model. There really aren't any creases or points of high curvature. As a result it is much less likely that an edge-based system could construct a representation either of the image or the model that would be stable under changes in illumination. Secondly, the albedo of the actual object is not exactly constant. The face contains eyebrows, lips and other regions where the albedo is not the same. As a result this is a test of EMMA's ability to handle objects where the assumption of constant albedo is violated. Thirdly, not all of the occluding contours of the object are present in the model. The model is truncated both at the chin and the forehead. As a result experiments with this model demonstrate that EMMA can work even when the occluding contours of the image and model are not in agreement.

In the previous experiment projecting points from the model into the image was sufficient to describe the model pose. Since the head model is very smooth and some occluding contours are missing simply projecting the model points into the image is not sufficient to determine the quality of an alignment. For our experiments with the head model we will display the original image, augmented with model points, alongside a rendered image of the model. Figures 5.11 and 5.12 show the model before and after alignment. In this experiment the model has been rotated 30 degrees around the vertical and translated 40 millimeters to the

[3]Thanks to Ron Kikinis for providing the Cyberware scan and for allowing me to take the images of him.

Figure 5.11: An initial incorrect pose. The model has been rotated 30 degrees about the vertical and translated 40 millimeters to the right. On the left is an image of the head along with a collection of points projected from the model. On the right is a rendering of the model in the same pose.

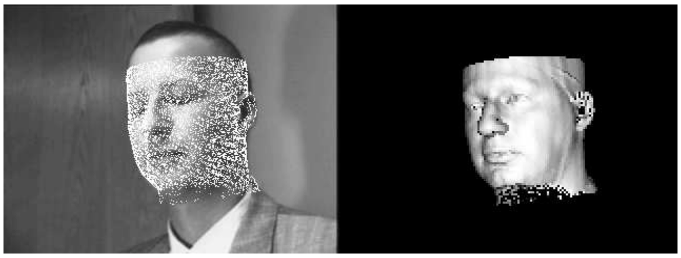


Figure 5.12: The final aligned poses. On the left is an image of the head along with a collection of points projected from the model. On the right is a rendering of the model in the same pose.

right. Figures 5.13 and 5.14 show another experiment where EMMA alignment corrects for a 150 millimeter translation in depth.

We have also tested EMMA alignment on a video sequence digitized from a video tape. The sequence was taken at the same time as the other images, though the camera and the lens were different. Ten frames were acquired from a video tape at 3 frames per second. The quality of the resulting images is very low. The images were degraded both by their storage on video tape and by the frame grabber that was used. It was somewhat surprising that these images worked nearly as well as the higher quality still frames.

Motion in the video sequence was tracked by sequentially aligning the model to each of the frames. The starting pose for each frame was obtained by using the final estimated pose from the previous frame. The starting pose for the first frame was hand selected so that EMMA alignment could acquire a good initial alignment. The sequence and pose estimates are displayed in Figure 5.15.

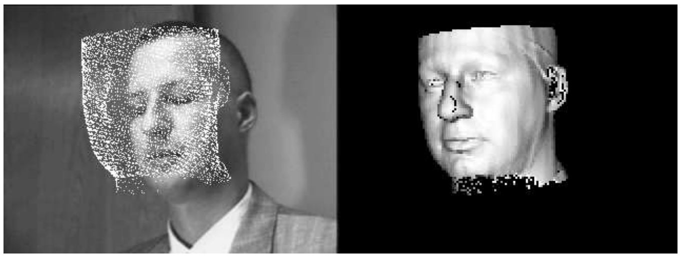


Figure 5.13: An initial incorrect pose. The model has been moved 150 millimeters toward the camera.

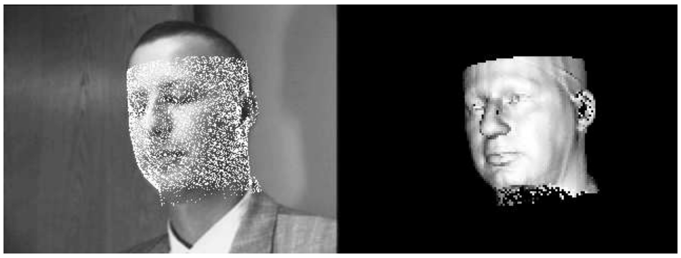


Figure 5.14: The final aligned poses.

Figure 5.15: Ten frames from a video sequence of RON's head.

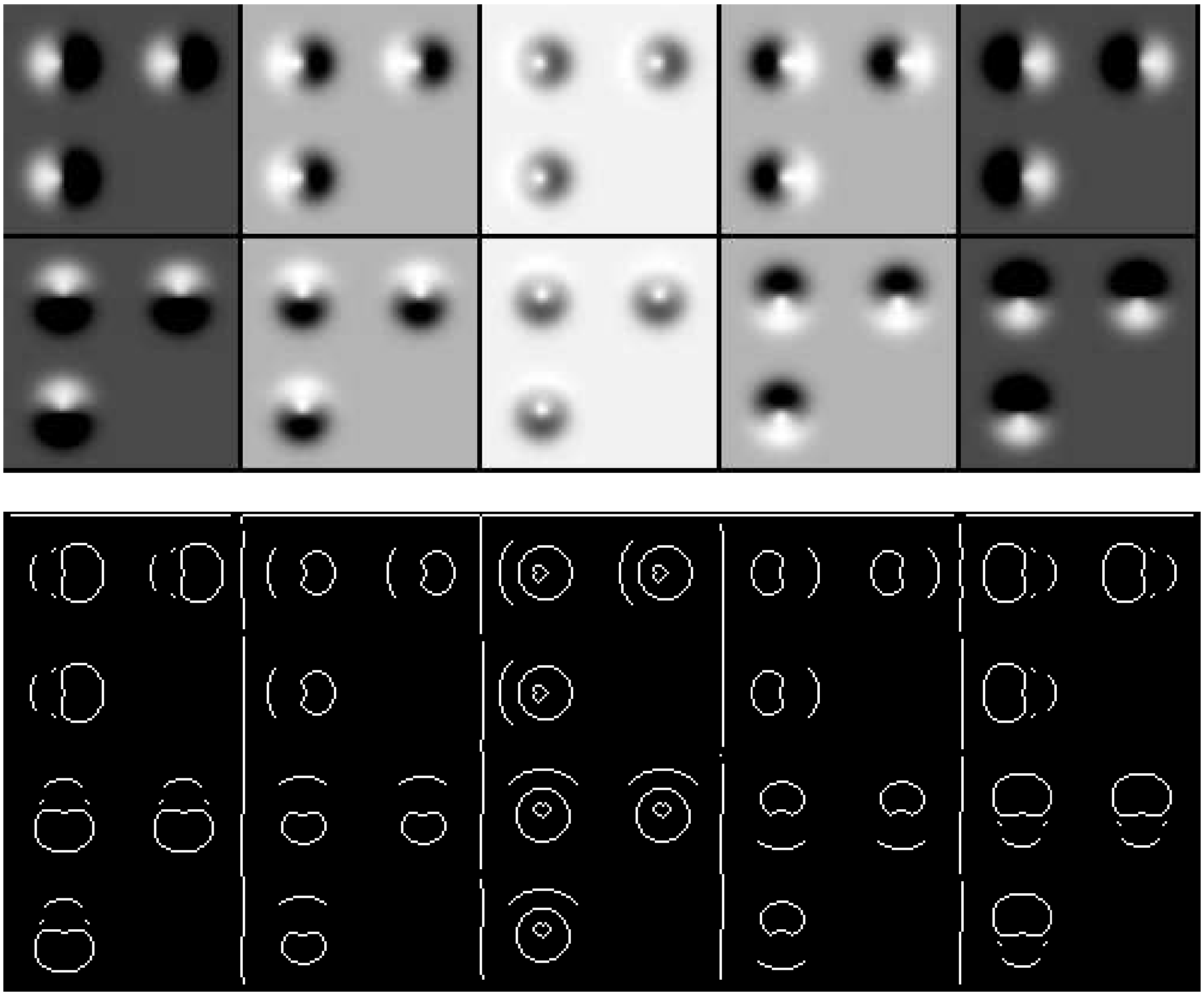


Figure 5.16: Bumps with Different Lighting, and their Edges

5.1.3 Alignment of Curved Surfaces

The third experiment is designed to explore the nature of the information that EMMA alignment uses to detect the correct pose. The previous two experiments, because they are based on real data, can be difficult to analyze. We would like to determine which component of the information in the image and model is critical to alignment. For example, it could be the case that EMMA alignment relies implicitly on intensity edges to match model to image. Or, it could be the case that the occluding contours of the object are of critical importance.

For this experiment we created a very simple, almost pathological, synthetic example.

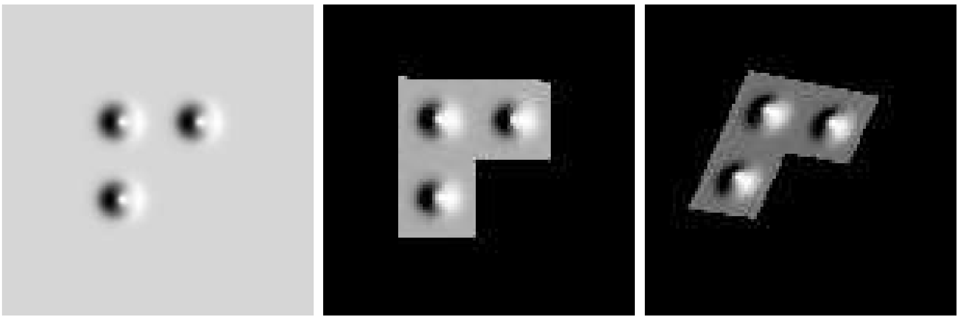


Figure 5.17: Target Image, Final Model Pose, and Initial Model Pose

The object is a set of three Gaussian shaped bumps in a flat patch of surface. This object in has no sharp edges and does not have a well defined occluding contour. Figure 5.16 shows ten different images of this object. Each image uses the same Lambertian reflectance model but has different illumination. Across the top row the light source moves gradually from left to right. In the second row the light source moves from top to bottom. Even for a simple Lambertian surface, image variation can be significant. Below this we show the output of a Canny edge detector run on the ten different images (Canny, 1986). The variation between the different edges extracted is quite striking.

Figure 5.17 shows the target image on the left. Here the bumps are inserted into an infinite plane. Also shown is a rendered version of a typical final and initial pose of the model. As in other experiments the rendered images of the model are made using a particular surface and lighting scheme; they are for visualization only and play no role in the alignment process. The black regions of the rendered image lie outside the borders of the model. Notice that the model's boundaries do not coincide with any discontinuities in the target image. Since there are neither stable edges nor usable occluding boundaries, we can conclude that EMMA alignment can proceed using *only* shading information.

The bumps each have a sigma of 7mm (the bump is about 3 sigma or 21mm wide). The bump height is 20mm. Lying together in the same plane they take up an L shaped region that is 100mm by 100mm. The true pose is 1000mm away from the camera and perpendicular to the camera axis. The camera has a viewing angle of 18 degrees. This experiment proceeds exactly as in the previous three dimensional experiments.

As we did with the skull model we performed an analysis of the reliability of the maxima of mutual information. These experiments summarized in Table 5.4.

$\triangle$T			$\triangle\theta$	INITIAL				FINAL				SUCCESS
X	Y	Z		σ_X	σ_Y	σ_Z	$\overline{\mid\triangle\theta\mid}$	σ_X	σ_Y	σ_Z	$\overline{\mid\triangle\theta\mid}$	
±mm			°	mm			°	mm			°	%
10	10	25	30	5.21	5.82	19.97	15.41	.56	.45	9.73	5.68	100
15	15	25	20	8.75	8.95	14.61	9.15	.82	.39	8.95	3.23	100
20	20	25	20	11.72	11.07	13.85	9.22	.65	.38	8.35	3.12	80

Table 5.4: Curved Surface Alignment Data

5.2 Medical Registration Experiments

EMMA alignment of three dimensional objects relies on the fact that the image is a function of the model and the lighting. It is not necessary that we know the exact nature of this function. In medical imaging we are faced with a different though related task. We are given two different observations of the same object. For example we may be given a *Computed Tomography* (CT) scan and *Magnetic Resonance Image* (MRI) of the same patient. Two scans are often obtained because neither gives perfect information about the patient. CT is good at finding bone. MRI is good for distinguishing soft tissue. These measurements are taken at different times with different machines. A clinician that would like to have information about bone and soft tissue must integrate the two scans into a single self-consistent picture. Once this is done the spatial relationships between structures in the two different scans become apparent. For example the distance between a tumor and a bone can be measured.

In Chapter 4 we argued that EMMA alignment should be able to align two signals whenever there is mutual information between them. In this experiment two different MR images of the same head will be aligned (see Figure 5.18). These images comprise the proton density and T2-weighted images of a double-echo MR scan. We have chosen these two MR scans for two reasons: (1) it is clear that the two images share a great deal of information, while they are not identical; and (2) since they are taken simultaneously the correct alignment should be close to the identity transformation. Because we know ground truth, we can evaluate the accuracy of the EMMA alignment procedure.

A typical initial alignment appears in Figure 5.19. Notice that this image is a scaled, sheared, rotated and translated version of the original. The final alignment is displayed as a checkerboard. Here every other 20x20 pixel block is taken either from the model image or aligned target image. Notice that the boundary of the brain in the two images is in close agreement.

We represent transformation space as a 6 element affine matrix that is used to project two dimensional points from the image into the model. This scheme can represent any combination of scaling, shearing, rotation and translation. The remaining algorithmic details are summarized in Table 5.5.

In order to determine the reliability and precision of the alignment procedure, 50 random-

1.	Define the model and image $u()$ and $v()$: $u()$ is the model image, $v()$ is the target image.
2.	Sampling x: Sample the pixels of the model image uniformly.
3.	Transformation space T: The space of affine transformations mapping pixels locations from the model into pixel locations in the image.
4.	Definition of $dv(y)/dy$: This is the intensity gradient.
5.	Distance metric: Euclidean distance.
6.	Variance, ψ: Assuming diagonal covariance matrices, three different variance are necessary: two for the joint entropy and one for the image entropy. The variances were 0.1 in all cases.
7.	Minimum probability, p_{min}: 0.01.
8.	Number of samples: One sample of 20 using cross-validation.
9.	Update rate, λ: 0.02 for 500 steps and then 0.005 for 500 steps.

Table 5.5: Summary of MRI alignment experiments.

ized alignments were performed. The initial transformations were randomly selected, having a translation of up to 35 pixels (this is about one third of the width of the head), a rotation of up to 30 degrees, and a scaling of up to 20%. The correct alignment was obtained in 100% of the experiments. After alignment the affine transformations had an average translation error of 0.1 pixels. The remaining affine parameters represent a mixing of rotation, scale and shearing. They are somewhat more difficult to interpret. Given that the correct transformation is the identity matrix we can evaluate the final matrices by measuring the difference from the identity. On average the coefficients of the final transformation were in error by 0.02. These experiments demonstrate that EMMA alignment is both precise and reliable.

These two MRI images are fairly similar. Good alignment could probably have been obtained with a normalized correlation metric. Normalized correlation assumes, at least locally, that one signal is a scaled and offset version of the other. Our technique makes no such assumption. In fact, it will work across a wide variety of non-linear transformations. More difficult alignment problems are easily simulated. In Figure 5.20 we show the model image after a non-monotonic non-linear function has been applied. Recall that initially the image lies in the range [0,1]. We subtract 0.5, square the result, and renormalize to [0,1]. This operation is shown at the right of the figure. After applying this non-linear transformation the two images are anti-correlated; no variant of correlation can correctly align them. EMMA

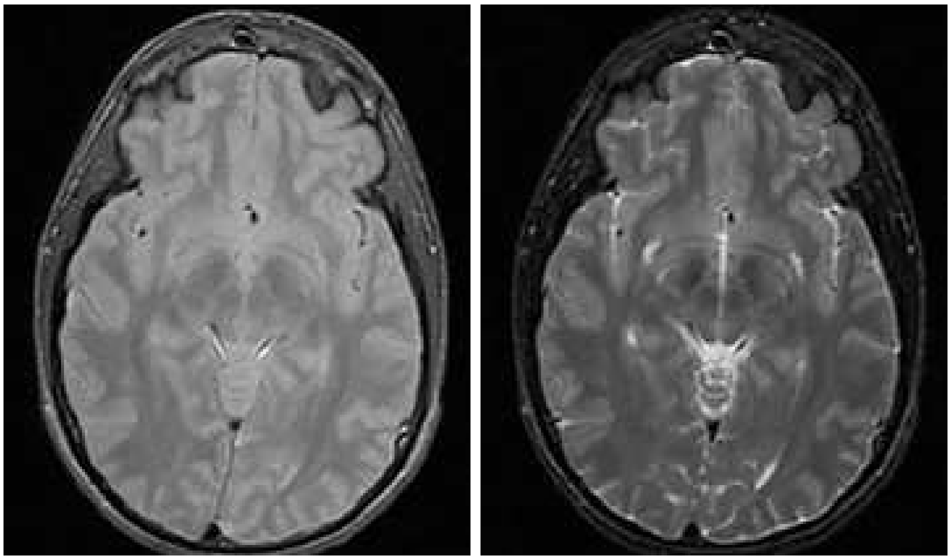


Figure 5.18: MR Images

alignment performance, however, is not effected.

5.2.1 Three Dimensional MR/CT Alignment

The medical alignment procedure described above can be extended to volumetric data. In the resulting system both the model and image are 3D arrays and a full three dimensional aligning transformation is estimated. Recently a number of three dimensional CT/MRI alignments have been performed. Because these results are preliminary, many of the experimental details are still in flux. Our description here will be necessarily brief. A more complete description can be found in (Wells III and Viola, 1995).

The scans used were obtained from the same patient at different times[4]. Display of these three dimensional scans, and their alignment, is a difficult problem. Though the entire scan cannot be shown, some feeling for the data can obtained be displaying the three central slices. The central slices are perpendicular planes that pass trough the centroid of the data. Figure 5.21 shows the three central slices of the CT scan. Figure 5.22 shows the initial

[4]The images and the standard transformation(s) were provided as part of the project, "Evaluation of Retrospective Image Registration", National Institutes of Health, Project Number 1 R01 NS33926-01, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN.

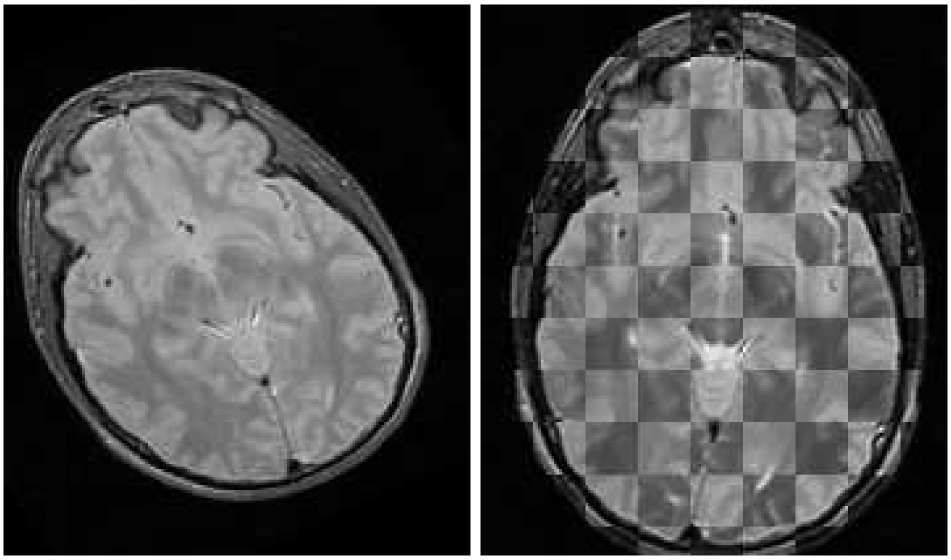


Figure 5.19: Initial Pose and Display of Result

alignment of the CT/MR pair as a checkerboard. When the signals involved are very different, the checkerboard representation can be somewhat confusing. Figure 5.23 shows the final alignment as the composition of the MR data with the intensity edges computed from the CT data. Notice the close agreement between the skull in both scans.

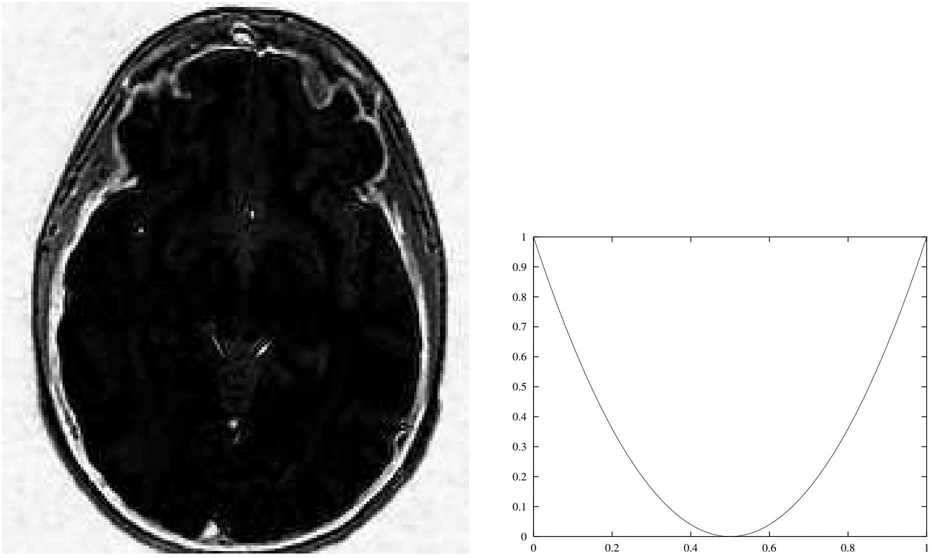


Figure 5.20: Transformed Model and the Transformation

Figure 5.21: The three central slices of the CT data used in the MR-CT experiments.

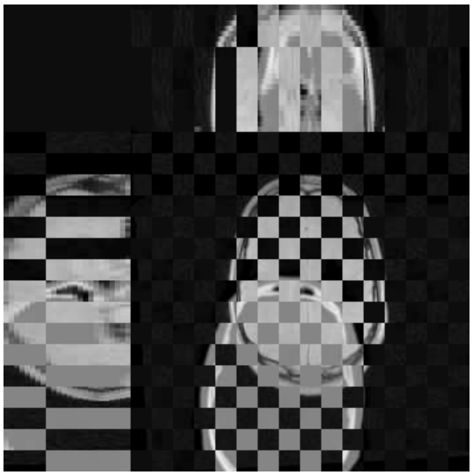


Figure 5.22: An initial condition for MR-CT registration by maximization of mutual information displayed as a checkerboard composite of the three central slices.

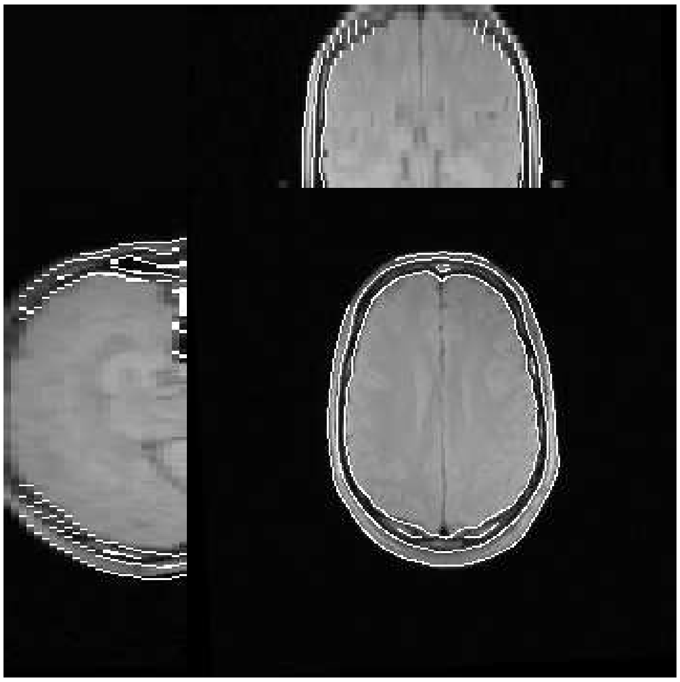


Figure 5.23: A final configuration for MR-CT registration by maximization of mutual information. The three central slices of the MRI data are shown with the edges from the registered CT data overlaid.

5.3 View Based Recognition Experiments

In the previous vision experiments we used knowledge of the physics of imaging to show that the surface normal of an object should be predictive of the intensity observed in an image. Unfortunately, in many experimental situations no three dimensional model is available. In these situations it is frequently the case that the only available information about an object is a collection of images taken under a variety conditions. One approach for solving problems like this is to use a collection of images as the model. This is often called a "view based" approach since the model is made up of a number of views of the model object. Given a novel image of some object, each model image is compared to it in turn. If some model image is "close enough" to the novel image, the model and novel image are considered aligned (or recognized). One can significantly reduce the number of model images required by adding an affine transformation to the comparison process. The novel image is then compared to each model image under a set of affine transformations. The most commonly used comparison metric is correlation. As we saw in Section 4.1.1, correlation makes the assumption that the model and the image are identical (or possibly related by linear function).

In general the set of images that can arise from a single object under varying illumination is very broad. Figure 5.24 shows two images of the same object in the same pose. These images are very different and are in fact anti-correlated: bright pixels in the left image correspond to dark pixels in the right image; dark pixels in the left image correspond to bright pixels in the right image. No variant of correlation could match these images together.

We have presented techniques based on entropy that can match both correlated and anti-correlated signals. These techniques require only that there is some consistent relationship between model and image. Discouragingly, it is not difficult to find two images of the same object for which there is no consistent relationship. Figure 5.25 shows a novel image which is aligned with the two model images. Figure 5.26 contains two scatter plots of the pixel values in the novel image versus the pixel values in the model images. Clearly, there is no simple consistent relationship displayed in either of these graphs. EMMA could not be used to match this novel image to either model image.

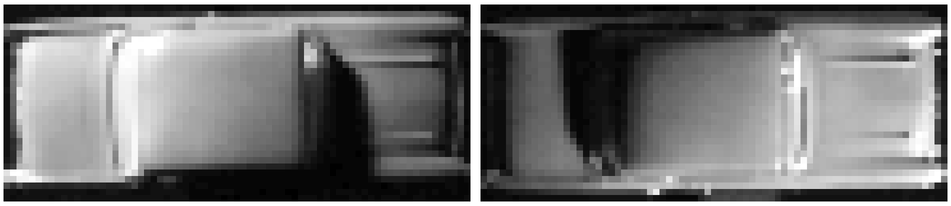


Figure 5.24: Car Model Images

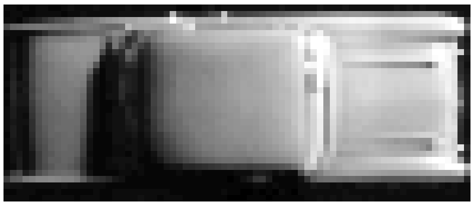


Figure 5.25: A novel image of the car model.

5.3.1 Photometric Stereo

By itself each model image does not contain enough information to constrain the match between image and model. However, it is well known that taken together a collection of images can be used to determine the 3D shape of an object. As we've seen the 3D shape is sufficient to constrain the match between image and model.

When multiple images of an object are available a technique called *photometric stereo* can be used to estimate its 3D shape (Horn, 1986). Photometric stereo works with images which are taken from the same location but under different illumination conditions. It is assumed that detailed information both about illumination and surface properties are available for each image. As a result a reflectance map can be computed for each image. The reflectance map determines the relationship between the normals of an object and the intensities in an image.

The reflectance map together with the intensity of a pixel acts as a constraint on the normal vector visible from that pixel. The allowable normals usually lie along a closed curve on the unit circle. From a second image, and its associated reflectance map, another set of allowable normals can be computed. By intersecting these constraints, two images are sufficient to determine the surface normal at each pixel. From the normals the shape can be obtained through integration.

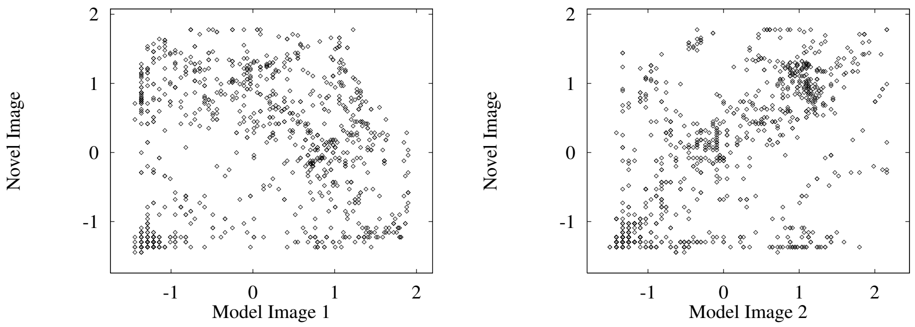


Figure 5.26: The relationship between pixels in the novel image and each of the model images.

Once the shape of the object is determined, the correct alignment could be computed using the three dimensional version of EMMA. The imaging function of this new two stage process is:

$$I(T(x_i)) = F(G(u_1(x_i), r_1, u_2(x_i), r_2), q)$$

where $G()$ is the photometric stereo function that takes two images and two reflectance maps and returns the shape, and $F()$ is our original imaging function which predicts image intensities from object normals.

In practice, however, performing photometric stereo requires the kind of detailed metric information about illumination that is only available under very controlled circumstances. One cannot use natural images where the lighting is unknown and difficult to determine. Luckily, we need not actually know $G()$, r_1, r_2, $F()$, or q. As long as they exist there will be high mutual information between any novel image and a pair of model images. This is the essence of view based EMMA alignment. We don't actually perform photometric stereo, we simply assume that it is possible. As a result a pair of images should give information about any third image.

To demonstrate this approach we have built a model using the two images in Figure 5.24. Figure 5.27 shows the target image, and the final pose obtained after alignment. Figure 5.28 shows the initial pose of the model.

Technically this experiment is very similar to the MRI alignment experiment. The main difference is that the model is constructed from a pair of model images. A sample of the model $u(x) = [u_1(x), u_2(x)]^T$ is a two dimensional vector containing the intensity of the two images at location x. This is similar to the two component representation of normal used in

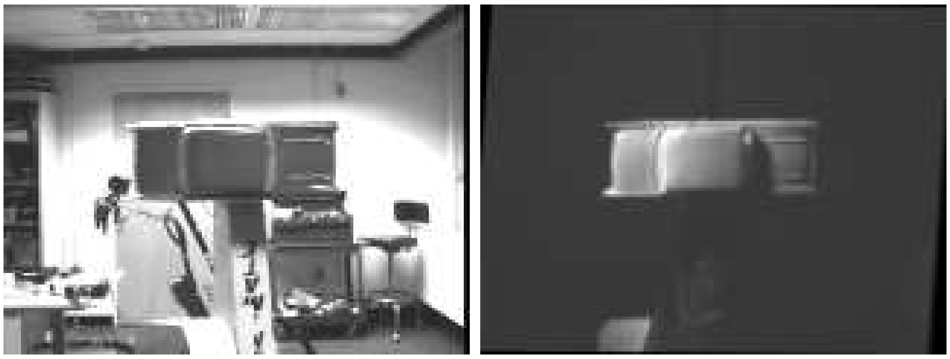


Figure 5.27: Car Image and Final Pose

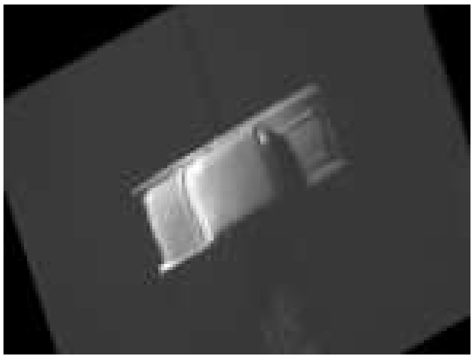


Figure 5.28: Initial Pose of Car Model

the three dimensional alignment experiments. For this experiment σ is 0.1. The parameters were updated for 1000 iterations at a rate of 0.002. From a set of randomized experiments we have determined that the capture range of the alignment procedure is about 40% of the length and width of the car, and 35 degrees of rotation.

5.4 Limitations of EMMA Alignment

Before we complete our discussion of EMMA alignment, several important caveats must be emphasized. EMMA alignment is *not* a recognition procedure. Though EMMA could well play a role in recognition there are two major missing components. The first missing component is an indexing scheme. EMMA alignment only works when the initial hypothetical pose is "close" to the true pose. A number of experiments have been performed in which an empirical estimate of "close" is determined (see Tables 5.3 and 5.4). The capture range of alignment is not large enough to expect that a randomly chosen transformation will converge to the true solution. As a result, some sort of additional procedure is required that can rapidly propose possible poses. Such a procedure is typically called an "indexing" scheme.

The second missing component is the recognition process itself. Recognition requires that a decision be made about whether the model object is really present in the image. Object recognition is very similar in concept to the problem of pattern classification (Duda and Hart, 1973; Fukunaga, 1990). Pattern classification is a process by which a novel input pattern is classified as an example of a class. The task is difficult when the structure of the class is complex and under-specified. For example, determining which stocks are a good investment is a classification task that even the most complex classifiers, human investors, have difficulty performing.

Pattern classification can be formulated as a maximum likelihood or maximum a posteriori problem. Given a novel pattern, the likelihood of each possible class is evaluated in turn. If one class is much more likely then any other, then that class is considered the correct class. Object recognition is a similar process. Given a novel image, the likelihood of each object model is evaluated. In order that there be confidence in the classification, it is particularly important that the most likely model be more likely than the null hypothesis: that the image does not contain any model. The likelihood of the null hypothesis is proportional to the unconditioned likelihood of an image.

As we have seen, log likelihood is closely related to entropy. As a result, EMMA can be used to define a classification procedure: the mutual information between each model and the novel image is evaluated; the model selected is the one that provides the most information about the image. What is missing is a reliable measure of the unconditioned entropy of an image (i.e. the information that the null hypothesis gives us about the image). The assumption that the pixels are independent underlies the EMMA estimate of image entropy. Though it leads to inaccuracy, the independence assumption has proven sufficient for alignment. This is primarily because alignment is a relative procedure. The model is adjusted so that the image is best explained. Recognition, because it is an absolute procedure, is not so forgiving. A more accurate estimate of image entropy will be required before EMMA can be used for object recognition.

Chapter 6

Other Applications of EMMA

The theory and algorithms presented in this thesis are quite general and can in principle be applied to a variety of problems. This chapter is devoted to a description of two such problems and their solutions. In the first part of the chapter we will show how EMMA can be used to correct images that have been have been "corrupted" by a slowly varying bias field. Examples include: MRI corruption that arises from non-uniformity in magnetic field, and lightness correction in visual images. The second part of the chapter is devoted to an application of stochastic gradient descent outside of entropy manipulation. Jones and Poggio have presented a system that aligns line drawings of faces with novel line drawings (Jones and Poggio, 1995). Their published work uses a complex second order gradient descent technique known as Levenberg-Marquardt. We will show that similar if not better results can be obtained with stochastic gradient descent. The resulting algorithm operates roughly 30 times as fast as the original.

6.1 Bias Compensation

A magnetic resonance image (MRI) is a 2 or 3 dimensional image that records the density of tissues inside the body. In the head, as in other parts of the body, there are a number of distinct tissue classes including: bone, water, white matter, grey matter, and fat. In principle the distribution of pixel values in an MRI scan should be clustered, with one cluster for each tissue class. In reality MRI signals are corrupted by a bias field, an additive or multiplicative offset that varies slowly in space. The bias field results from unavoidable variations in magnetic field (see (Wells III et al., 1994) for an overview of this problem). The bias field makes constructing automatic tissue classifiers difficult.

Wells et al. have built a system for bias correction around the assumption that an uncorrupted MRI scan will have a particular distribution of pixel values. This distribution will have a peak for each type of tissue. Using an explicit physical model of MRI image formation they construct a prior model for this distribution as a mixture of Gaussians, with one Gaussian for each tissue type. The model can then be used to compute the likelihood of an MRI. Corrupted MRI's will be unlikely because the bias field blurs together the clusters. Wells et al. use maximum likelihood to select the correction field (the inverse of the bias field) that makes a corrupted MRI most likely.

To reiterate, their system finds an estimated correction field that when applied to the data makes it look like a particular type of clustered multi-class data. Applying the correction field sharpens up the classes and makes automatic classification easy. As in the learning problems encountered in previous chapters, some assumption about the nature of the correction field is necessary to condition the problem. If we have prior knowledge that the bias field varies slowly across space, the correction field should also vary slowly. Wells et al. assume that the bias field is smooth. To encourage smoothness they introduce a probabilistic prior in which smooth fields are more likely than non-smooth ones.

The main disadvantage of their MRI correction system is that it requires a fairly accurate model of the tissue distribution. These models can be difficult to construct. Furthermore, since the model includes estimates for the relative proportions of the tissue types, a different model is required for each region of the body.

Using entropy we can proceed in a much less model-based way. Since Wells et al.'s technique has proven to be quite effective, we can safely assume that the pixel values of an uncorrupted MRI image are clustered into distinct classes. Such a distribution should have low entropy. Corruption from the bias field blurs together the clusters. The bias field acts like noise, adding entropy to the pixel distribution. This insight is the central idea behind our approach. We attempt to find the low-frequency correction field that when applied to the image, makes the pixel distribution have a lower entropy. The resulting "bias corrected" image will have a tighter clustering than the original distribution.

Insuring the smoothness of the correction field can be tricky. Wells et al. estimate a dense correction field, with one estimate for every pixel in the MRI. They insure smoothness by periodically, at every iteration, smoothing the correction field estimates. Another approach would be to represent the bias field as smooth in the first place. This can be done parametrically by representing the correction field as a smooth parameterized function. Or it can be done using a low-frequency correction image, with say 1 pixel for every 10 in the image. Another approach, one which is guaranteed to be better when it is possible, is to represent the correction field with an explicit physical model. In the case of MRI, physics can be used to show that the bias field should be a low order polynomial of location (M. Tincher and Williams, 1993). We take this approach for correcting MRI scans, representing the correction field as a third degree polynomial in the x and y coordinates of the scan.

The code that minimizes the entropy of the MRI is very similar to the other entropy manipulation applications we have described. Once again we sample points from the image, x, where each point now has a value, $v(x)$, and a current estimate for the bias field, $b(x)$. Proceeding as we have before, we approximate the entropy of the bias compensated image, $c(x) = v(x) - b(x)$ (see Equation (3.22)). The bias field is adjusted to minimize $h^*(c(x))$ by taking steps in the direction of the derivative as approximated in Equation 3.24. In this case the parameters over which we are minimizing entropy are the coefficients of the bias field polynomial. The derivatives of the bias field, since they are polynomial, are easy to compute.

The first experiment is identical to a synthetic experiment proposed by Wells et al. A binary checkerboard is used as a prototypical example of a two class image. Half of the pixels belong to the black class, the other half to the white class. The pixel entropy of a checkerboard is very low. The checkerboard is then corrupted by a large unknown bias field (see Figures 6.1 and 6.2). The corruption is so large that any cluster structure in the data has disappeared. This is apparent both from the distribution of pixels and a thresholding of the corrupted image (see Figures 6.3 and 6.2). Entropy minimization comes very close to exactly compensating for the bias field. Figure 6.2 shows the corrected image. Its distribution is also shown in Figure 6.3. The image is not perfectly corrected because we use a different bias field representation than Wells et al.

One of the goals of the work by Wells et al. is to correctly classify white versus grey matter in the brain (see (Bezdek et al., 1993) for a comprehensive overview of MRI segmentation). They show that classification is much easier if the bias field of a scan is known beforehand.

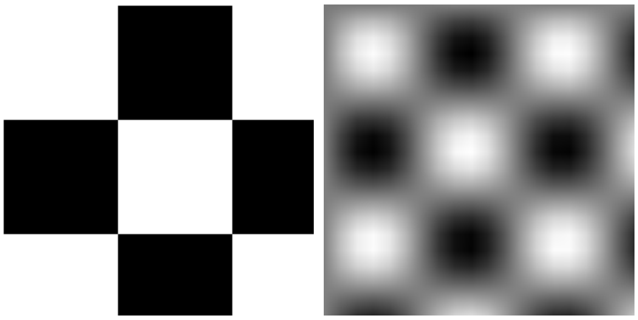


Figure 6.1: The original checkerboard image and the bias field that corrupts it.

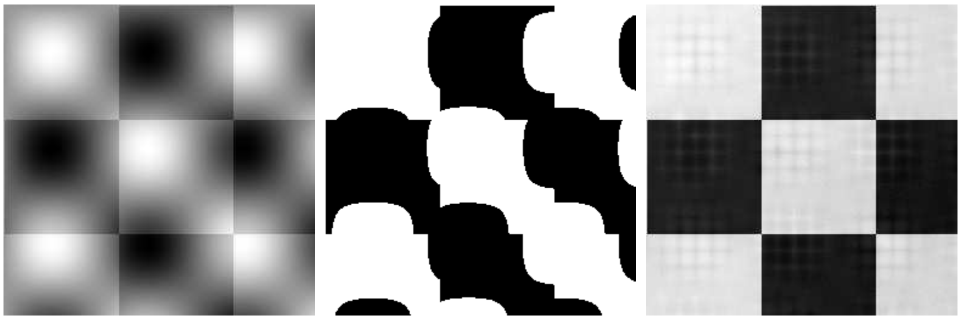


Figure 6.2: Left: the corrupted checkerboard. Center: a thresholded version of the corrupted image. Right: the corrected checkerboard.

We have performed a number of experiments where EMMA is used to find the correction field. While Wells et al.'s system is designed to give a tissue classification for the corrected scan, EMMA based correction does not provide a classification. Instead we will examine the distribution of the corrected scan and attempt to determine if the scan has been corrected in a way that would make classification of white and grey matter easier. Figure 6.4 shows a slice of MRI data taken from a brain. Figure 6.5 shows the histogram of the scan before and after correction. In the histogram of the original scan white and grey matter tissue classes are confounded into a single peak ranging from about 0.4 to 0.7. The histogram of the corrected scan shows much better separation between these two classes. We can highlight the pixels in this distribution by mapping all the pixels below the first peak to black, all the pixels above the second peak to white, and linearly scaling between (white matter appears darker than grey matter in this MRI scan). Figure 6.6 shows the original and corrected scans in this manner. Notice that the inhomogeneity in the original image becomes immediately apparent. The lower left hand portion of the original scan is dark. Since the corrected scan does not show this inhomogeneity, the white and grey matter of the corrected scan are distinct.

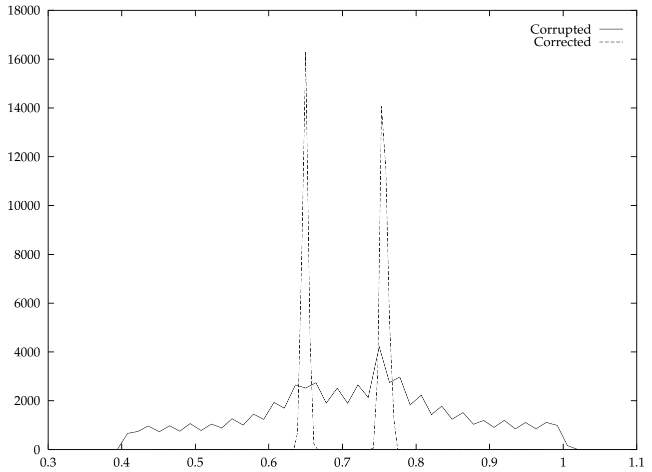


Figure 6.3: The pixel density of the corrupted checkerboard and the compensated checkerboard.

Figure 6.4: A slice from an MRI scan of a head.

A Second Experiment

The procedure for bias field correction has been repeated for a number of different scans. Figures 6.7, 6.8 and 6.9 show the results of an experiment performed on a coronal slice of an MRI scan.

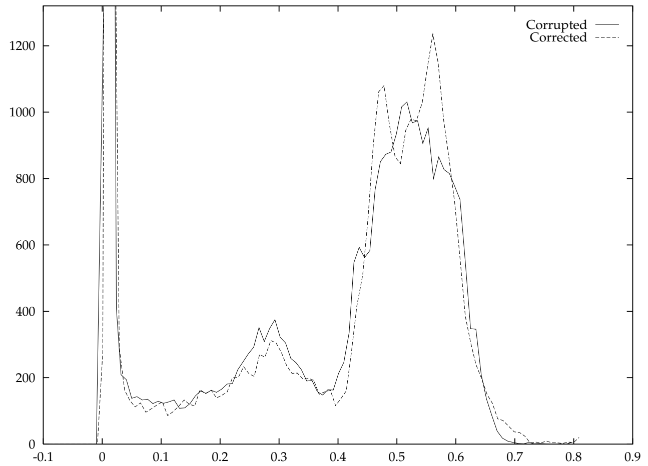


Figure 6.5: The distribution of pixel value in the MRI scan before and after correction.

Experimental Details

All of the experiments use a smoothing function variance of 0.01. The learning rate was -0.003. The sample size was 30. Every run was concluded after 8000 parameter updates. The p_{min} value was 0.01.

The checkerboard image is 256 by 256 pixels. The checkerboard experiment uses a bias field which is a 20 by 20 pixel low resolution image. This 20 by 20 image is then bilinearly interpolated to create a continuous bias offset for each pixel in the checkerboard. The 400 parameters in this bias field are the most ever simultaneously approximated by an EMMA based technique.

The MRI scans varied in size from 100 square to 256 square. The bias field for the MRI experiments was a 3rd order polynomial in x and y location.

6.2 Alignment of Line Drawings

Jones and Poggio have constructed a system that automatically analyzes hand drawn faces (Jones and Poggio, 1995). Their system can estimate the emotional content of these faces (i.e. it can determine that you have drawn a very happy face that is somewhat surprised). Their

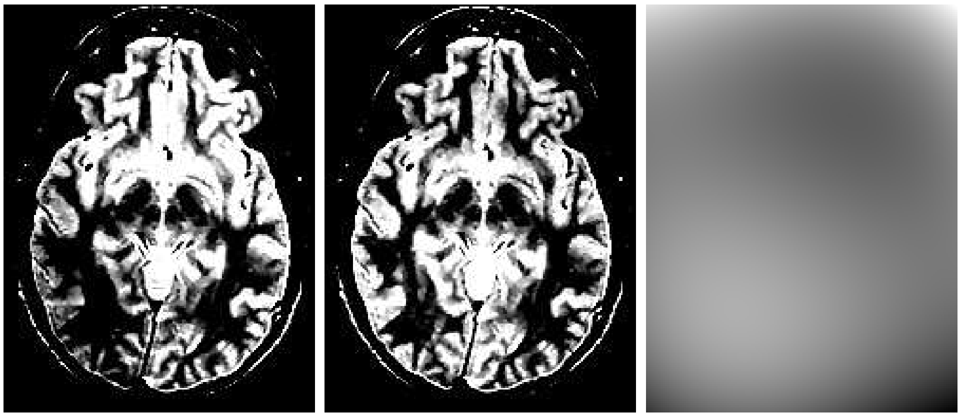


Figure 6.6: At left is the MRI scan shown with a modified intensity scale. The inhomogeneity is not much more apparent. In the center is the correct scan shown using the same intensity scale. At right is the estimated correction field.

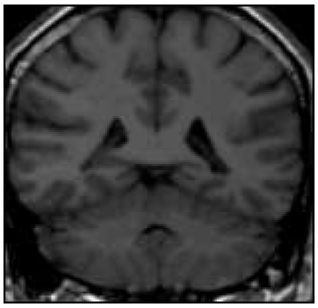


Figure 6.7: A coronal slice from an MRI scan of a head.

system works by constructing a non-rigid transformation that maps a novel drawing onto a hand drawn "neutral" face (see figure Figure 6.10). The shape of the non-rigid transformation determines the emotion of the face.

Jones and Poggio use a representation for non-rigid transformations that is called *flow*. A flow is an image of displacement vectors. They search for a flow that minimizes the difference between the base image $b(x)$ and the novel image $n(x)$,

$$C(f) = \sum (n(x) - b(x + f(x)))^2 \quad . \tag{6.1}$$

where f is the flow image and the summation is over all of the pixels in the novel image. The problem of non-rigidly transforming one image into another has been very heavily studied in computer vision. In most previous work flow is represented directly as an image of displace-

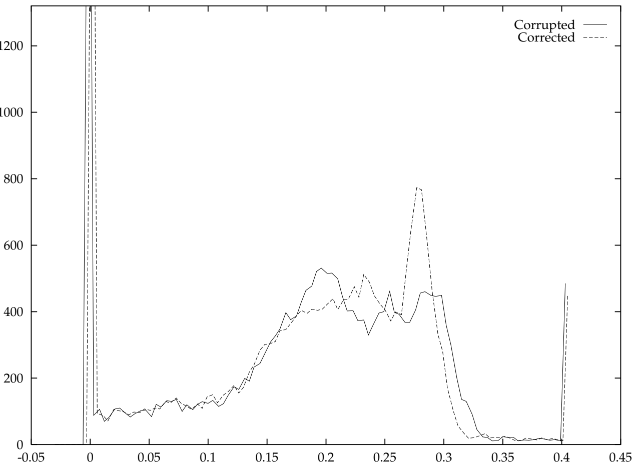


Figure 6.8: The distribution of pixel value in the MRI scan before and after correction. Solid line is the original distribution. The dashed line is the corrected distribution.

Figure 6.9: The same scan but with a modified intensity scale. Points above the intensity for grey matter appear white. Points below the intensity for white matter appear black. There is a linear scale between. Notice that the lower part of the image is darker in the uncorrected image. On the right is the estimated bias.

ment vectors. The search problem is then conditioned with the addition of a smoothness prior over flow images. Jones and Poggio decompose flow into a linear combination of component flows,

$$C(\{\alpha_i\}) = \sum \left(n(x) - b(x + \sum_i \alpha_i f_i(x)) \right)^2 \quad . \tag{6.2}$$

The search for flow then becomes an unconstrained optimization over the parameters α_i.

Each component flow represents a different type of emotion. For example, one kind of flow might transform a neutral face into a smiling face. Another might make a face look more surprised. These flows can be mixed together to produce images that have combination

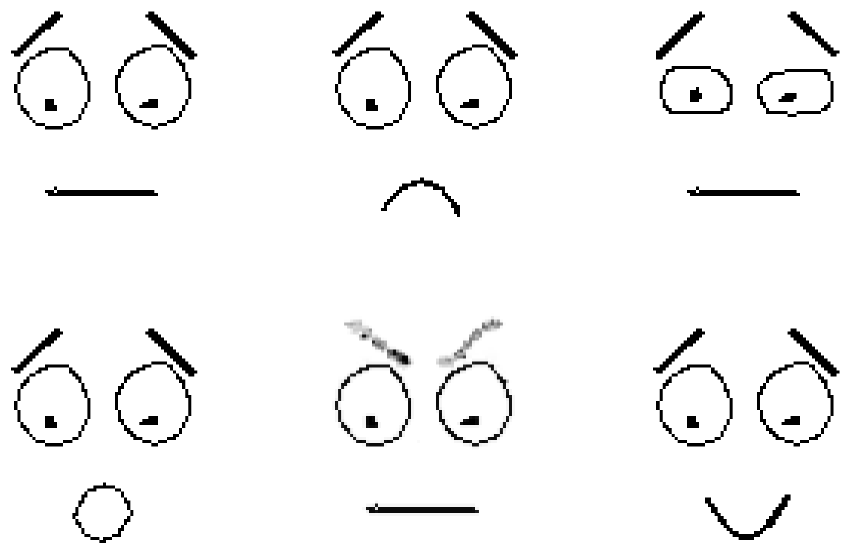


Figure 6.10: The first face is the "neutral" face. The others are "frown", "narrow eyes", "surprise", "eyebrows", and "smile".

of properties. Given a novel image, a set of α_i's are determined that provide for the closest match with the neutral face. These α_i's determine to what extent the face is smiling or frowning. Figure 6.11 shows several novel images and the best reconstruction obtained by transforming the neutral face.

In their paper Jones and Poggio use Levenberg-Marquardt, a second order gradient descent procedure, to determine the flow parameters (see (William H. Press and Veterling, 1992) for an excellent discussion optimization techniques). Together we have replaced this technique with a much simpler stochastic gradient descent procedure. Over the course of many experiments stochastic alignment has improved running times from 60 seconds to 2 seconds. The quality of the minima found with stochastic gradient descent is equivalent to, if not better, than Levenberg-Marquardt. Moreover, stochastic gradient descent rarely if ever gets stuck in local minima. There were a number of cases where Levenberg-Marquardt converges far from the the best solution. On these same problems, stochastic gradient descent almost always finds a good solution.

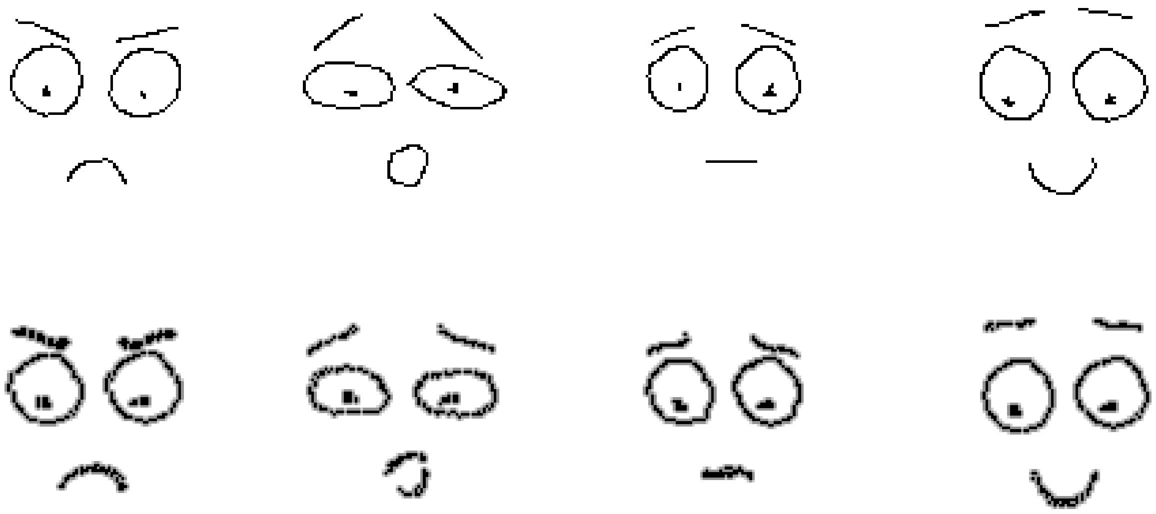


Figure 6.11: A series of novel faces with the best reconstruction displayed below.

Chapter 7

Conclusion

Maximization of mutual information appears to be a new and powerful means of performing local alignment of objects and images. In a typical vision application it is an intensity-based, rather than feature based method. While intensity based, it is more robust than traditional correlation, as shown by the insensitivity to lighting demonstrated in the experiment of Section 5.1.3. In addition, the method is insensitive to negating the image data, as well as a variety of non-linear transformations, which would defeat conventional intensity-based correlation.

The weaknesses of intensity correlation may be corrected, to some extent, by performing correlations on the magnitude of the brightness gradient. This, as well as edge-based matching techniques, can perform well on objects having discontinuous surface properties, or useful silhouettes. These approaches work because the image counterparts of these discontinuities are reasonably stable with respect to illumination.

Gradient magnitude correlation, as well as edge-based methods can have serious difficulties in domains lacking discontinuities, such as the example shown in Section 5.1.3, because neither edges, nor their precursor, gradient magnitude, are stable in image position with respect to lighting changes (see Figure 5.16). While our technique works well using only shading, it also works well in domains having surface property discontinuities and silhouette information (see Section 5.1.1).

7.1 Related Work

Alignment by extremizing properties of the joint signal has been used by Hill and Hawkes (Hill et al., 1994) to align MRI, CT, and other medical image modalities. They use third order moments to characterize the clustering of the joint data. We believe that mutual information is perhaps a more direct measure of the salient property of the joint data at alignment, and demonstrate an efficient means of estimating and extremizing it.

There are many schemes that represent models and images by collections of edges and define a distance metric between them that is proportional to the number of edges that coincide (see the excellent survey articles: (Besl and Jain, 1985; Chin and Dyer, 1986)). A smooth, optimizable version of this metric can be defined by introducing a penalty both for unmatched edges and for the distance between those that are matched (Lowe, 1985; Wells III, 1992b; Huttenlocher et al., 1991). This metric can then be used both for image/model comparison and for pose refinement. Edge based metrics can work under a variety of different lighting conditions, but they make two very strong assumptions: the edges that arise are stable under changes in lighting, and the models are well described as a collection of edges. Clearly smoothly curved objects are a real problem for these techniques. As we alluded before, Wells has performed a number of experiments where he attempts to match edges that are extracted under varying lighting. In general for even moderately curved objects, the number of unstable and therefore unreliable edges is problematic. Faces, cars, fruit and a myriad of other objects have proven to be very difficult to model using edges.

Others use more direct techniques to build models. Generally these approaches revolve around the use of the image itself as an object model. Objects need not have edges to be well represented in this way, but care must be taken to deal with changes in lighting and pose. Turk and Pentland have used a large collection of face images to train a system to construct representations that are invariant to some changes in lighting and pose (Turk and Pentland, 1991). These representations are a projection onto the largest eigenvectors of the distribution of images within the collection. Their system addresses the problem of recognition rather than alignment, and as a result much of the emphasis and many of the results are different. For instance, it is not clear how much variation in pose can be handled by their system. We do not see a straightforward extension of this or similar eigenspace work to the problem of pose refinement. On a related note Shashua has shown that all of the images, under different lighting, of a Lambertian surface are a linear combination of any *three* of the images

(Shashua, 1992). This also bears a clear relation to the work of Turk and Pentland in that the eigenvectors of a set of images of an object should span this three dimensional space.

Entropy is playing an ever increasing role within the field of neural networks. We know of no work on the alignment of models and images, but there has been work using entropy and information in vision problems. None of these techniques uses a non-parametric scheme for density/entropy estimation as we do. In most cases the distributions are assumed to be either binomial or Gaussian. This both simplifies and limits such approaches.

Linsker has used the concept of information maximization to motivate a theory of development in the primary visual cortex (Linsker, 1986). He has been able to predict the development of receptive fields that are very reminiscent of the ones found in the primate visual cortex. He uses a Gaussian model both for the signal and the noise.

Becker and Hinton have used the maximization of mutual information as a framework for learning different low-level processing algorithms such as disparity estimation and curvature estimation (Becker and Hinton, 1992). They assume that the signals whose mutual information is to be maximized are Gaussian. In addition, they assume that the only joint information between images is the information that they wish to extract (i.e. they train their disparity detectors on random dot stereograms).

Finally, Bell has used a measure of information to separate signals that have been linearly mixed together (Bell and Sejnowski, 1995). His technique assumes that the different mixed signals carry little mutual information. While he does not assume that the distribution has a particular functional form, he does assume that the distribution is well matched to a preselected transfer function. For example, a Gaussian is well matched to the logistic function because applying a correctly positioned and scaled logistic function results in a uniform distribution.

7.2 A Parallel with Geometrical Alignment

EMMA bears some similarity to methods used for evaluating and adjusting *geometrical* alignment. These similarities may be seen by revisiting the entropy derivative of Equation 3.28, and comparing it to the derivative of the following construct.

We define $\mathcal{D}$, half the averaged Mahalonobis distance between values in B and their nearest correspondences in A,

$$\mathcal{D}(T) \equiv \frac{1}{N_B} \sum_{z_i \in B} \min_{z_j \in A} \frac{1}{2} D(z_i - z_j) \ . \tag{7.1}$$

Locally (away from discontinuities), the derivative of the above expression is

$$\frac{d}{dT}\mathcal{D}(T) \equiv \frac{1}{N_B} \sum_{z_i \in B} \min_{z_j \in A} \frac{d}{dT}(\frac{1}{2} D(z_i - z_j)) \ .$$

Comparing the above expression with Equation 3.28, we see the following analogy. If the transformation T is adjusted to reduce the averaged squared "differences" between points in B and their counterparts from A that are nearest in *signal value*, then a reduction in entropy is obtained. This is intuitive, in that entropy will be lower if clusters in "signal value" are tighter so that nearby signal differences will be smaller. The approximation of this analogy is due to the dissimilarity between max and softmax.

Equation 7.1 is essentially the measure used in chamfer matching techniques, such as the method described by Borgefors (Borgefors, 1988). Huttenlocher (Huttenlocher et al., 1991) has used a related measure in feature matching applications, the Hausdorff distance, which uses maximum instead of the sum that appears in Equation 7.1. The similarity between geometrical matching and entropy becomes even stronger if one uses the softmax operation to weight the closest element rather than simply selecting the closest, as Wells has (Wells III, 1992b; Wells III, 1992a).

We reiterate that in vision applications, these methods have typically been used to measure aggregate geometrical distance, while here we are measuring aggregate distances among signal values (typically intensities, brightnesses, or surface properties).

Appendix A

Appendix

A.1 Gradient Descent

In a number problems described in this thesis one must find a set of parameters that extremizes an evaluation function. Examples include: (1) finding the parameters of density so that the likelihood of sample is maximized; (2) finding the pose parameters that align a model and an image best; and (3) finding the weights of a neural network so that it approximates a function best. In each case there is a function of a parameters set $F(p)$, whose value is to be either maximized or minimized. The parameters are continuous variables, and we are therefore faced with an infinite number of possible solutions. The gradient descent procedure is an effective though greedy technique for searching such a space.

There are many closely related gradient descent algorithms. Here we will describe the simplest: steepest descent or hill climbing. Starting from an initial guess for the parameters, steepest descent is an iterative procedure that uses the partial derivatives of a function to construct an improved estimate for its parameters. Each parameter is updated by

$$p \leftarrow p + \lambda \frac{\partial F(p)}{\partial p} \quad .$$

The update rate λ (which is also known as the learning rate) must be chosen carefully. When

λ is sufficiently small one can use a Taylor expansion of $F()$ to prove that

$$F(p + \lambda \frac{\partial F(p)}{\partial p}) \geq F(p) \ .$$

When λ is too small p might take arbitrarily long to approach a maximum. If λ is chosen correctly p will converge toward the maximum relatively rapidly.

There are many gradient based techniques that attempt to speed the rate of convergence of p. Second order techniques such as Levenberg-Marquart and Newton-Raphson use the second derivatives of $F(p)$ to re-estimate λ. Conjugate gradient techniques attempt to find better directions than the gradient of $F()$. In every case one must be careful that the theoretical advantages of the algorithm are not outweighed by the costs of computing it. Researchers in neural networks have found that for many problems it is difficult to realize any actual improvement in convergence speed. The problems for which steepest descent works as well as more complex techniques include functions where there are a large number of parameters—this makes computing the second derivatives quite expensive.

Bibliography

Anderson, J. and Rosenfeld, E., editors (1988). *Neurocomputing: Foundations of Research.* MIT Press, Cambridge.

Baclawski, K., Rota, G.-C., and Billey, S. (1990). Introduction to the theory of probability. MIT course notes for 18.313.

Becker, S. and Hinton, G. E. (1992). Learning to make coherent predictions in domains with discontinuities. In Moody, J. E., Hanson, S. J., and Lippmann, R. P., editors, *Advances in Neural Information Processing*, volume 4, Denver 1991. Morgan Kaufmann, San Mateo.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximisation approach to blind separation. In *Advances in Neural Information Processing*, volume 7, Denver 1994. Morgan Kaufmann, San Francisco.

Besl, P. and Jain, R. (1985). Three-Dimensional Object Recognition. *Computing Surveys*, 17:75–145.

Bezdek, J., Hall, L., and Clarke, L. (1993). Review of MR Image Segmentation Techniques using Pattern Recognition. *Medical Physics*, 20(4):1033 – 1048.

Bienenstock, E., Cooper, L., and Munro, P. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2. Reprinted in (Anderson and Rosenfeld, 1988).

Borgefors, G. (1988). Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm. *IEEE Transactions PAMI*, 10(6):849–865.

Bridle, J. S. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In Touretzky, D. S., editor, *Advances in Neural Information Processing 2*, pages 211–217. Morgan Kaufman.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions PAMI*, PAMI-8(6):679–698.

Chin, R. and Dyer, C. (1986). Model-Based Recognition in Robot Vision. *Computing Surveys*, 18:67–108.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley and Sons.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1 – 38.

Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis*. John Wiley and Sons.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Harcourt Brace Jovanovich.

Grimson, W., Lozano-Pérez, T., Wells, W., et al. (1994). An Automatic Registration Method for Frameless Stereotaxy, Image Guided Surgery, and Enhanced Realigy Visualization. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, WA. IEEE.

Haykin, S. (1994). *Neural Networks: A comprehensive foundation*. Macmillan College Publishing.

Hill, D. L., Studholme, C., and Hawkes, D. J. (1994). Voxel Similarity Measures for Automated Image Registration. In *Proceedings of the Third Conference on Visualization in Biomedical Computing*, pages 205 – 216. SPIE.

Horn, B. (1986). *Robot Vision*. McGraw-Hill, New York.

Huttenlocher, D., Kedem, K., Sharir, K., and Sharir, M. (1991). The Upper Envelope of Voronoi Surfaces and its Applications. In *Proceedings of the Seventh ACM Symposium on Computational Geometry*, pages 194–293.

Intrator, N. and Cooper, L. N. (1992). Objective function formulation of the bcm theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17.

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Jones, M. and Poggio, T. (1995). Model-based matching of line drawings by linear combinations of prototypes. *Proceedings of the International Conference on Computer Vision*.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by Simulated Annealing. *Science*, 220(4598):671–680.

Leclerc, Y. (1988). Constructing simple stable descriptions for image partitioning. *Proceedings Image Understanding Workshop*, 1:365–382.

Linsker, R. (1986). From basic network principles to neural architecture. *Proceedings of the National Academy of Sciences, USA*, 83:7508–7512, 8390–8394, 8779–8783.

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer*, pages 105–117.

Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press.

Lowe, D. (1985). *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers.

M. Tincher, C. R. Meyer, R. G. and Williams, D. M. (1993). Polynomial modeling and reduction of rf body coil spatial inhomogeneity in mri. *IEEE Trans. Med. Imaging*, 12(2):361–365.

Mellor, J. (1995). Realtime camera calibration for enhanced reality visualization. In *Computer Vision, Virtual Reality and Robotics in Medicine*, pages 471–475. Nice, France.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., third edition.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Robbins, H. and Munroe, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Schraudolph, N. N. and Sejnowski, T. J. (1993). Unsupervised discrimination of clustered data via optimization of binary information gain. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing*, volume 5, pages 499–506, Denver 1992. Morgan Kaufmann, San Mateo.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 and 623–656.

Shashua, A. (1992). *Geometry and Photometry in 3D Visual Recognition*. PhD thesis, M.I.T Artificial Intelligence Laboratory, AI-TR-1401.

Turk, M. and Pentland, A. (1991). Face Recognition using Eigenfaces. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, Lahaina, Maui, Hawaii. IEEE.

Wells III, W. (1992a). Posterior Marginal Pose Estimation. In *Proceedings: Image Understanding Workshop*, pages 745 – 751. Morgan Kaufmann.

Wells III, W. (1992b). *Statistical Object Recognition*. PhD thesis, MIT Department Electrical Engineering and Computer Science, Cambridge, Mass. MIT AI Laboratory TR 1398.

Wells III, W., Grimson, W., Kikinis, R., and Jolesz, F. (1994). Statistical Gain Correction and Segmentation of MRI Data. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, Wash. IEEE , *Submitted*.

Wells III, W. M. and Viola, P. A. (1995). Multi-modal volume registration by maximization of mutual information. In preparation.

Widrow, B. and Hoff, M. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pages 96–104. IRE, New York.

William H. Press, Brian P. Flannery, S. A. T. and Veterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, England, second edition edition.