Alignment by Maximization of Mutual Information

Emanuil Hristov

February 24, 2020

I will present you a simplified version of an already existing paper(It's in the same repository under the name "Original Paper"). I will try to present this paper so it can be available to a much bigger auditory of people, mainly those who aren't very familiar with further math and for this paper mainly calculus 1. I recommend everyone who think has bigger interest in this paper and/or has finished this "simplified" version to check the original paper. I will explain the mathematical and physical concepts used in the paper, assuming a High School mathematics level. I will break down the paper and explain each of its parts. I recommend not rushing trough this paper and taking the time to understand each part, even if this means rereading it again and again. All recommendations and questions are welcome.

Part I
Introduction

I want to remind again that this is just a simplified version of an already existing paper.

The Introduction is going to be pretty much the same as in the original paper, due to the lack of complicated mathematics and the importance of this part for the general understanding of the paper.

The core problem, as we can see in the title of the paper, is alignment and to be precise - alignment of model and image data in object recognition and image registration. For example in medical imaging data from one type of sensor must be aligned with that from another and this can be very difficult task. In order to overcome this difficulty this paper present a theoretical approach, which is making few assumption about the nature of the imaging process.

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. The predicted image can then be compared to the actual image directly. If the object model and pose are correct the predicted and actual images should 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 $v(x)$ and an image $v(y)$ we can formulate an imaging equation,

$$v(T(x)) = F(v(x), q) + \eta \qquad (1)$$

or equivalently,

$$v(y) = F(u(T^{-1}(y), q) + \eta \qquad (2)$$

The imaging equation has three 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,$F(v(x), q)$. 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. Finally, η is a random variable that models noise in the imaging process. In most cases the noise is assumed Gaussian.
Alignment can be a difficult problem for a number of reasons:

- F , the imaging function of the physical world, can be difficult to model.
- q, the exogenous parameters, are not necessarily known and can be difficult to find. For example computing the lighting in an image is a non-trivial problem.
- T , 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. 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 v and v, there could be no meaningful F . We propose to finesse the problem of finding and computing F and q by dealing with this mutual information directly. We will present an algorithm that aligns by maximizing the mutual information between model and image.

This paper will present a new approach for evaluating entropy and mutual information called EMMA [1]. It is distinguished in two ways: 1) EMMA does not require a prior model for the functional form of the distribution of the data; 2) entropy can be maximized (or minimized) efficiently using stochastic approximation.

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

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

data from one type of sensor (such as magnetic resonance imaging—MRI) must be aligned to data from another sensor (such as computed tomography—CT).

Part II
Alignment Example

In this part we shell look at an example of alignment in order to get a better understanding of the theory. Before we continue I shell explain a couple of mathematical concepts. If you are familiar with them you can skip this explanation.

- Summation - Denoted with $\sum$ and it is the addition of a sequence of any kinds of numbers. General the notation is defined as $\sum_{i=m}^{n} a_i = a_m + a_{m+1} + a_{m+2} + ... + a_{n-1} + a_n$.
- Lambertian surface - The brightness of a Lambertian surface appears uniform from any direction of view. In other words, the luminance of the surface is isotropic. Lambertian surfaces are often referred to as ideal diffusion surfaces. [2]
- Lambert's law - 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.
- Probability density function (PDF) - "a statistical expression that defines a probability distribution (the likelihood of an outcome) for a discrete random variable (e.g., a stock or ETF) as opposed to a continuous random variable. The difference between a discrete random variable is that you can identify an exact value of the variable. For instance, the value for the variable, e.g., a stock price, only goes two decimal points beyond the decimal (e.g. 52.55), while a continuous variable could have an infinite number of values (e.g. 52.5572389658...)." [3] In other words if you want to get the probability of getting 5 you have to look at the probability of getting number between, for example, 4.9999 and 5.0001, and this probability will be equal to the area under the graph between those two points.
- Probability distribution - is a function that provides us the probabilities of all possible outcomes of a stochastic process. It can be thought of, as a description of the stochastic process, in terms of the probabilities of events. The most commonly occurring distribution is the Gaussian Distribution or the Normal Distribution.[4]
- Normal distribution (Gaussian distribution) - This is a type of continuous probability distribution for a real-valued random variable.[5] The general form of its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

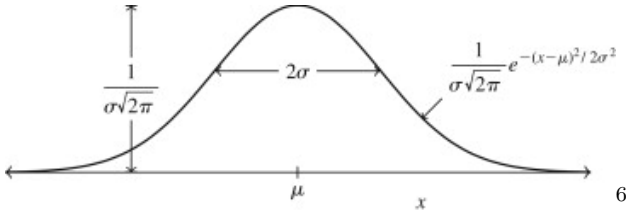
[2] Lambertian surface - link
[3] What is a PDF? - link
[4] Probability distribution - link
[5] Normal distribution - link

The parameter μ is the mean or expectation of the distribution (and also its median and mode) and σ is its standard deviation. The variance of the distribution is σ^2. A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

I will present another explanation too - The gaussian distribution is a means to measure the uncertainty of a variable that is continuous between $-\infty$ and $+\infty$. The distribution is centered at mean, μ. The width depends on the parameter σ, the standard deviation (variance, σ^2). Naturally, area under the curve equals 1. [4]he area under a graph is calculated by integrating

- Gaussian noise - the values that the noise can take on are Gaussian-distributed. The density function of Gaussian noise, q, with mean μ and variance σ^2 is

$$p_q(a) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$



6

We can finally begin with this part of the paper. One of the alignment problems that we will address involves finding the pose of a threedimensional 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 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: RK). A depth map is an image that displays the depth from the camera to every visible point on the object model.

Figure 2 contains two renderings of the object model. These synthetic images are constructed assuming that the 3D model has a Lambertian surface and that the model is illuminated from the right. It is almost immediately obvious that the model on the left of the figure is more closely aligned to the video image than the model on the right.

[4]T

Figure 1: Two different views of RK. On the left is a video image. On the right is a depth map of a model of RK that describes the distance to each of the visible points of the model. Closer points are rendered brighter than more distant ones.

Figure 2: At left is a rendering of a 3D model of RK. 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.

Lambert's law is perhaps the simplest model of surface reflectivity. It is an accurate model of the reflectance of a matte 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:

$$v(T(x)) = \sum_i \alpha_i \vec{l_i}.u(x) \qquad (3)$$

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

$$\upsilon(T(x)) = \sum_i R(\alpha_i, \vec{l_i}, \vec{o}, u(x)) \qquad (4)$$

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 both about alignment and image formation. 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 difficult to visualize.

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, for visible patches, is determined by two parameters: the x and y components. The image 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.

Figure 3 contains a graph of the intensities along a single scan-line of the image of RK. Figure 4 shows similar data for the correctly aligned model of RK. 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 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 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.

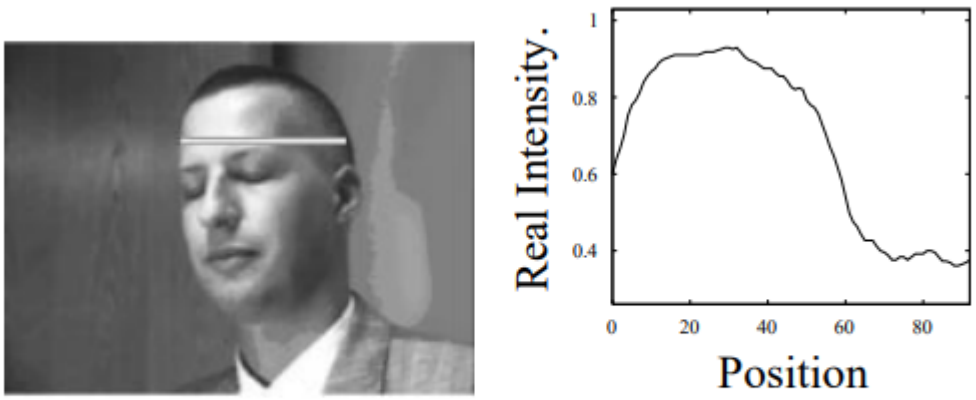


Figure 3: On the left is a video image of RK with the single scan-line highlighted. On the right is a graph of the intensities observed along this scan line.

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 generally lead to a consistent relationship. Conversely, when model and image are misaligned the relationship between intensity and normal is inconsistent.

Part III

A Formal Definition of Consistency

As in the precious part I will first explain the mathematical concepts used in this part.

- Products of sequences - Denoted with $\prod$ and it is the multiplication of a sequence of any kinds of number. General notation is defined as $\prod_{i=m}^{n} = x_m \cdot x_{m+1} \cdot x_{m+2} \cdot \ldots \cdot x_{n-1} \cdot x_n$
- Conditional probability - this is a measure of the probability of an even happening(occurring) given that other event has happened(occurred). If we have two events A and B, the probability of A happening given B is denoted by P(A|B). This can be defined as $\mathrm{P(A|B)} = \frac{P(A \bigcap B)}{P(B)}$, that is the probability of the joints of the events A and B, and the probability of B.
- Principle of Maximum Likelihood - This will be a longer one

The principle of maximum likelihood is a method of obtaining the optimum values of the parameters that define a model. And while doing so, you increase the likelihood of your model reaching the "true" model. We can see this in the following example: I am borrowing this amazing toy example from Nando de Fretais's lecture, to illustrate the principle of maximum likelihood. Consider 3 data points, $y_1 = 1, y_2 = 0.5, y_3 = 1.5$, which are independent and drawn from a guassian with unknown mean θ and variance 1. Let's say we have two choices for $\theta : (1, 2.5)$. Which would you choose? Which model (θ) would explain the data better?

In general, any data point drawn from a gaussian with mean θ and and variance 1, can be written as,

$$y_i \sim N(\theta, 1) = \theta + N(0, 1)$$

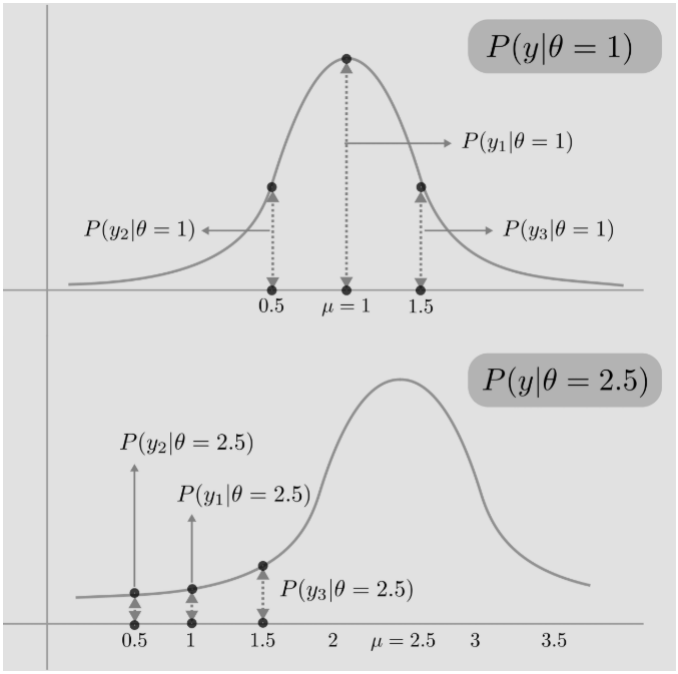
θ, the mean, shifts the center of the standard normal distribution ($\mu = 0$ and $\sigma^2 = 1$).

The likelihood of data (y_1, y_2, y_3) having been drawn from $N(\theta, 1)$, can be defined as,

$$P(y_1, y_2, y_3|\theta) = P(y_1|\theta)P(y_2|\theta)P(y_3|\theta)$$

as y_1, y_2, y_3 are independent.

Now, we have two normal distributions defined by $\theta = 1$ and $\theta = 2.5$. Let us draw both and plot the data points. In the figure below, notice the dotted lines that connect the bell curve to the data points. Consider the point $y_2 = 0.5$ in the first distribution ($N(\mu = 1, \sigma^2 = 1)$). The length of the dotted line gives the probability of the $y_2 = 0.5$ being drawn from $N(\mu = 1, \sigma^2 = 1)$. [7]



[7] Maximum likelyhood - link

The likelihood of data (y_1, y_2, y_3) having been drawn from $N(\mu = 1, \sigma^2 = 1)$, is given by,

$P(y_1, y_2, y_3|\theta) = P(y_1|\theta)P(y_2|\theta)P(y_3|\theta)$ The individual probabilities in the equation above, are equal to the heights of corresponding dotted lines in the figure. We see that the likelihood, given by the product of individual probabilities of data points given model, is basically the product of lengths of dotted lines. It is obvious that the likelihood of model $\theta = 1$ is higher. We choose the model $(\theta = 1)$, that maximizes the likelihood.[7]

Alignment can be performed by jointly searching over the space of possible imaging functions, exogenous parameters and transformations. The principle of maximum likelihood can be used to motivate this procedure. We can try to find the imaging function and exogenous variables that make the image most likely,

$$p(\upsilon|u, T) = \max_{F,q} \prod_{x_a} p(\eta = \upsilon(T(x_a, q))p(F)p(q)) \tag{6}$$

[7]

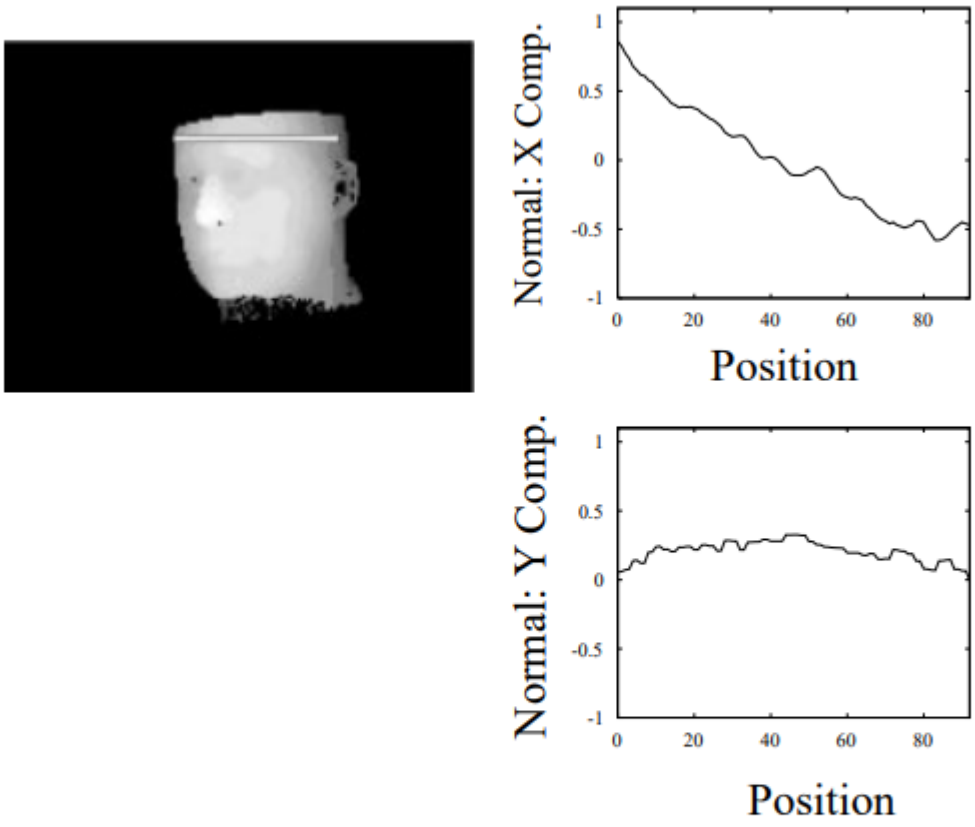


Figure 4: On the left is a depth map of RK with the single scan-line highlighted. At top right is a graph of the x component of the surface normal. On the bottom right is the y component of the normal.

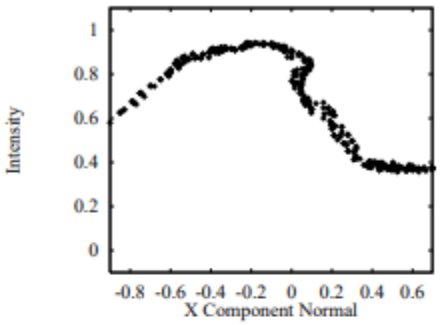


Figure 5: THE ALIGNED CASE: A scatter plot of the intensity of the video image versus the x component of the surface normal from the model. The image and model are correctly aligned.

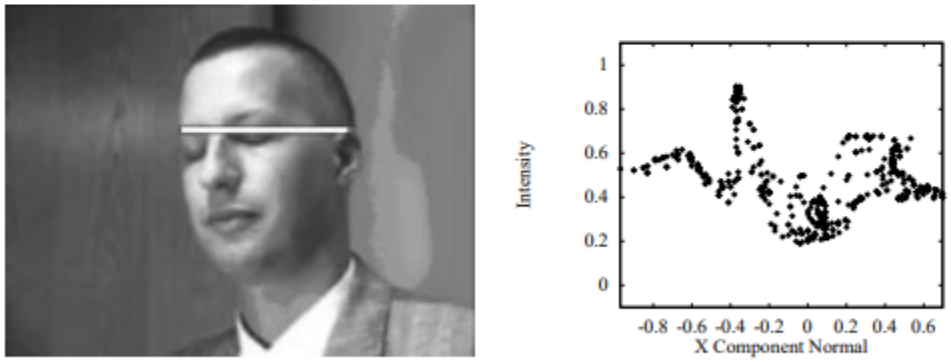


Figure 6: THE MISALIGNED CASE: On the left is the misaligned scan-line from the video image of RK. On the right is a scatter plot of the intensity of this part of the video image versus the x component of the surface normal from the model.

Using (6) we can define an alignment procedure as a nested search: i) given an estimate for the transformation, find F and q that make the image most likely; ii) 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 $v(T(x_a), u(x_a))$.

The search for F is not a simple process. The range of possible imaging functions is of course infinite. In order to condition the search it is necessary to make a set of assumptions about the form of F . In addition some assumptions about the smoothness of F are necessary to insure convergence of the nested search for the maximum of (6). These assumptions can be enforced by formulating 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. The combined function is best thought of as 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 6 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 $|\upsilon(T(x_a) - \upsilon(T(x_b)))|$. One ad-hoc technique for estimating consistency is to pick a similarity constant and evaluate the following sum:

$$Consistency(T) = -\sum_{x_a \neq x_b} g_\psi(u(x_b) - u(x_a))(\upsilon(T(x_b)) - \upsilon(T(x_a)))^2 \tag{7}$$

where g_ψ is a Gaussian with standard deviation ψ , and the sum is over points from the model, x_a and x_b . In order to minimize this measure, points that are close together in value must be more consistent, and those further apart less so.

An important drawback of consistency is that it 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. We now have two alternatives for alignment when the imaging function is unknown: a theoretical technique that may be intractable, and an outwardly efficient ad-hoc technique that has a number of important difficulties. One would like to find a technique that combines the best features from each approach. We propose that the complex search for the most likely imaging function, F_q , be replaced with a simpler search for the most consistent imaging function.

One type of function approximator that maximizes consistency is known as kernel regression or the weighted neighbor approximator:

$$F*(u, a) = \frac{\sum_{x_a} R(u - u(x_a))\upsilon(T(x_a))}{\sum_{x_a} R(u - u(x_a))} \tag{8}$$

The weighting function R usually has a maximum at zero, and falls off asymptotically away from zero. F* can be used to estimate the likelihood of a transformation as we did in (6). This formulation can be much more efficient than a naive implementation of (6) since there is no need to search for F_q . The model, image, and transformation define F* directly.

Figure 7 shows the weighted neighbor approximation to the data from the RK alignments (in these graphs R is the Gaussian density function with variance 0.0003). Notice $F*$ fits the aligned model much better than the misaligned model. Assuming that the noise is G the log likelihood of the aligned model, 1079.49, is much larger than the log likelihood of the misaligned model, 537.342^2

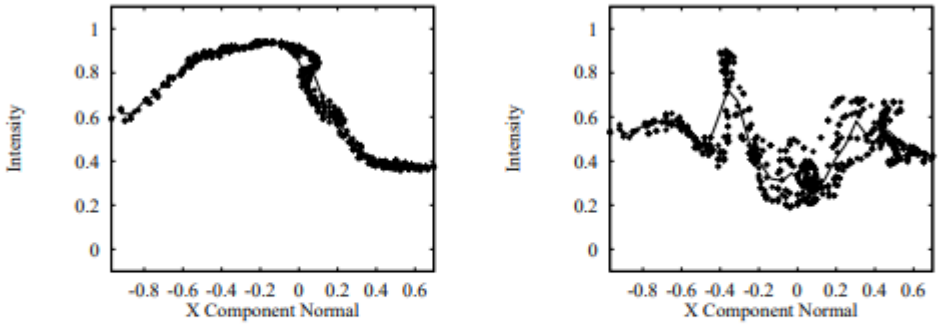


Figure 7: The joint distribution of data from the aligned and misaligned case above (left: aligned, right: misaligned). The weighted neighbor function approximation is show as a thin black line.

Part IV
From Likelihood to Entropy