Using Information Maximization for Alignment of Objects with Albedos

by

Andreas Argiriou

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Bachelor of Science in Computer Science and Engineering
and Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1997

Author .....
Department of Electrical Engineering and Computer Science
July 8, 1997

Certified by .....
Tomás Lozano-Pérez
Cecil H. Green Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Using Information Maximization for Alignment of Objects with Albedos

by

Andreas Argiriou

Submitted to the Department of Electrical Engineering and Computer Science
on July 8, 1997, in partial fulfillment of the
requirements for the degree of
Bachelor of Science in Computer Science and Engineering
and Master of Engineering in Electrical Engineering and Computer Science

Abstract

The problem of comparing and aligning three-dimensional models to images has attracted a significant amount of interest in the field of computer vision. However, it has turned out to be a complex and difficult task. Traditional approaches to alignment are plagued by a variety of pitfalls, which mainly stem from the complex nature of the imaging process. The image of an object is a function of many parameters, including the properties and position of the object as well as extraneous factors, like illumination conditions. The variations in illumination, in particular, have a strong effect on the image of an object but cannot be easily incorporated in an aligning algorithm.

A recent alignment method, however, tackles the problem with greater success using a radically different approach. This method offers a more appropriate measure for comparing such dissimilar entities as image intensities and object properties, based on information theory. Alignment is achieved by maximizing an estimate of the mutual information between image and model using a stochastic gradient procedure. We apply this method to alignment of some characteristic geometrical objects with albedos. We show that albedos can be incorporated in the information maximization scheme with very few modifications.

Thesis Supervisor: Tomás Lozano-Pérez
Title: Cecil H. Green Professor of Computer Science and Engineering

Acknowledgments

I wish to thank Professor Tomás Lozano-Pérez for his support and advice throughout the writing of this thesis. I am also grateful to William Wells for providing most of the code necessary for an aligning algorithm. A big thanks goes to both, and to Paul Viola, for answering questions I had or for offering hints and explanations. In addition, it would not be possible to omit several friends whose noble company I have enjoyed and to whom I am indebted: (dear racing car co-driver) Paris Smaragdis and Theo Evgeniou for their generous help; Elias Vyzas for all sorts of assistance; Latex/Matlab expert Carl Livadas; and language advisor Emanuela Binello. But most of all, I feel I should thank my mother and brother for their continuous support and tolerance.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

The problem of *alignment* is one of many examples demonstrating the difficulties of developing computational algorithms for vision. Alignment is the process of determining the rotation and translation under which an object appears in an image. Recovering or adjusting the attitude of an object in space, as alignment does, is required in many applications and is a general and important problem in itself. Furthermore, it is one of the basic tasks a competent vision system should perform.

In a possible situation of alignment, an image of an object and a model of some kind for the object are given. The task of aligning the model to the image means finding an appropriate transformation or *pose* (rotation and translation) that brings the model as close as possible to the object displayed in the image. Such a situation is illustrated in Figure 1-1. Clearly, there are a small range of poses that can be considered aligning poses and are satisfactory solutions to the problem. The eye can easily distinguish between the "correct" pose and the "incorrect" one of Figure 1-1. In fact, alignment is a routine task for the human visual system, as everyday experience can testify. However, the same task has turned out to be quite challenging for computers and programs have not yet managed to tackle it with complete success.

Traditional approaches to the problem often make use of some measure of comparison, based on edge-finding, correlation etc., that may give an indication of how well

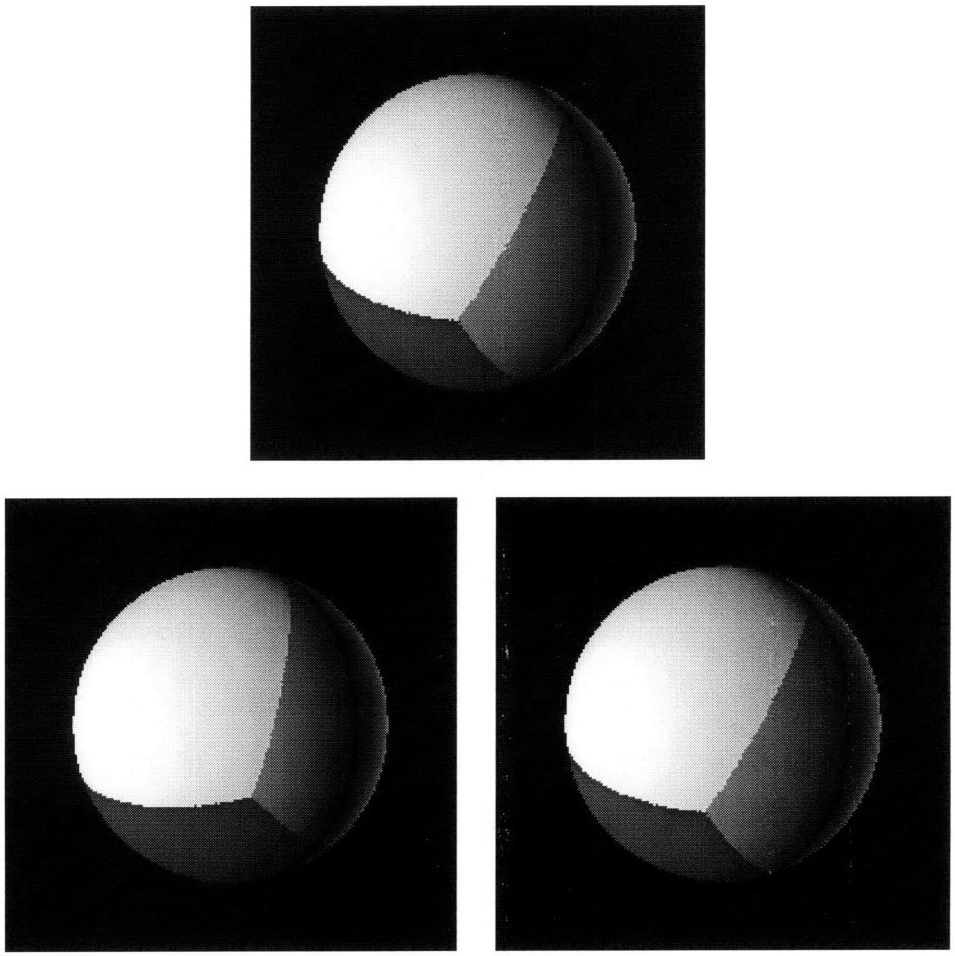

Figure 1-1: On top an image of a spherical object with a non-uniform albedo pattern. On bottom, two poses of this object are illustrated using the same imaging process. On the right is the aligning pose and on the left a non-aligning one.

model and image match. A significant drawback of these methods is their sensitivity to changes of lighting and their unreliability under non-constant imaging conditions. Such methods usually work under restrictive assumptions because they rely on relatively simple and low-level heuristics or measures. By contrast, the nature of the imaging process is generally unknown, complicated and depends on many parameters, some of which are hard to estimate. Consequently, such algorithms work well under certain conditions but poorly when the conditions change.

A recent development in this area of research employing information theoretic concepts seems to be more promising. Increasing interest in methods involving *entropy*

and *mutual information* has appeared lately in such areas as neural networks, ICA (Independent Component Analysis) and BSS (Blind Source Separation). One could briefly mention [5], [2], [1] as a first glimpse of the related literature. This research in turn has inspired a few new information theoretic approaches to vision, mainly in the work of Bell and Sejnowski ([3]) on feature detection and the work of Viola ([11]) on alignment.

In his work, Viola presents a generic estimate of the mutual information of two signals and then uses it for image-to-image alignment and alignment of three-dimensional models to images. The central idea is that the mutual information between the surface properties of the model and the image intensities is maximized when model and image are aligned. Compared to traditional approaches, mutual information turns out to provide a more robust and more reliable measure for matching models to images while requiring minimal a priori assumptions about the nature of the problem. It avoids the complications stemming from our ignorance of the imaging process, which has been the source of most of the weaknesses in other methods. Because mutual information is a generic statistical measure, it can work equally well under varying illuminating conditions and for a wide range of alignment situations that may have little in common with each other.

This new and promising information maximization method for alignment will be the subject of the present thesis. More specifically, this thesis focuses on the alignment of three-dimensional models to images for a few familiar and simple geometrical objects. Similar examples appear in [11], where the models of a skull and a human face are successfully aligned to real images. However, we will allow the possibility of varying albedos, which will enable us to consider objects with "painted" patterns on their surface. As will be shown, the information maximization procedure for alignment can be modified slightly to accommodate for the presence of albedos. After presenting the algorithm, we will investigate and interpret its behavior and finally we will evaluate its performance with some characteristic experiments.

1.2 Thesis Overview

The thesis consists of three main chapters. Chapter 2 presents the necessary background from probability theory and the theory of non-parametric estimation; it also includes some useful basic knowledge from the field of computer vision. Chapter 3 includes the formulation of alignment, a brief motivation for using mutual information and presents the derivations of the alignment algorithm. It also includes a discussion of information maximization illustrated by experiments. Finally, the actual alignment experiments and their results are presented in Chapter 4.

Chapter 2

Background

2.1 Concepts of Probability and Information Theory

In this section some basic definitions of probability and information theory are reviewed. Most of the discussion is based on the treatment of the subjects in [9] and [4].

Entropy and Mutual Information

There are several ways to characterize and compare probability distributions in accordance to some desirable criterion or another. One such characterization is provided by the computation of the *entropy* of a random variable, which is a measure of the "randomness" and "uncertainty" of a distribution. For a discrete random variable X, entropy is defined as the sum

$$H(X) = -\sum_{i=1}^{N} p_i \ln p_i \tag{2.1}$$

over the sample space of the random variable. In this definition, $p_i \ln p_i$ is considered to equal zero when $p_i = 0$, since $\lim_{x \to 0}(x \ln x) = 0$.

Combining the definition of expectation of a random variable with (2.1), we obtain

the alternative expression

$$H(X) = -E[\ln(p(X))]. \tag{2.2}$$

It can be shown that entropy is positive and bounded by $\ln N$:

$$0 \leq H(X) \leq \ln N. \tag{2.3}$$

The conditions for the lower bound are that all but one p_i are zero. This is the case of lowest uncertainty and randomness. In other words, drawing a sample from this distribution ensures that there will be no randomness in the outcome. Still another interpretation states that a distribution as simple as this has zero *complexity*.

The opposite observations are true for the upper bound case. This case occurs when the distribution is uniform, namely when $p_1 = \cdots = p_N = 1/N$. There is maximum uncertainty and randomness since no particular value is more probable than any other. Moreover, the distribution is more complex in the sense that there are no predominant values in it.

The definition of entropy can be generalized to the continuous domain by a process of approximating the continuous probability distribution by a discrete one of increasing sample space. The resulting definition is the following. Let X be a *continuous* random variable and $p(X)$ its density function. The *differential entropy* of X is then defined as the integral

$$h(X) = -\int p(x) \ln p(x)\, dx, \tag{2.4}$$

where the integration covers the sample space of X. Again, $p(x) \ln p(x)$ is considered to be equal to 0 whenever $p(x) = 0$. In the case that X is a random vector consisting of the random variables $X_1, X_2, \ldots, X_n$, Equation (2.4) defines what is also called the *joint entropy* of these variables.

Like discrete entropy, differential entropy is a measure of uncertainty, unpredictability and complexity. Higher entropies correspond to distributions which are

more "even", more "random" and more complex. On the other hand, less even and simpler distributions are characterized by lower entropies. Unlike discrete entropy, differential entropy can take both negative and positive real values.

In addition to the standard definition, it is often useful to express entropy as an expectation. The expected value of the random variable $g(X)$ formed by some function g can be given by the following theorem:

$$E[g(X)] = \int g(x)p(x)\,dx. \tag{2.5}$$

Applying this and (2.4), we find that

$$h(X) = E[-\ln p(X)]. \tag{2.6}$$

It can also be shown that the joint entropy of two *independent* random variables is the sum of their entropies. In fact,

$$h(X,Y) \leq h(X) + h(Y), \tag{2.7}$$

with the equality holding when X, Y are independent.

In coding theory, entropy has been viewed as a measure of *information* since it gives an expression for the expected minimal description length of a discrete distribution. There are also other widely used information theoretic measures, like *Kullback-Leibler entropy* and *mutual information*. The mutual information of two random variables, in particular, will be most relevant to our purposes. It is a measure of how much information each of the two variables gives about the other and is defined as the sum of the two entropies minus the joint entropy:

$$I(X,Y) = h(X) + h(Y) - h(X,Y). \tag{2.8}$$

The definition is similar for the discrete case:

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{2.9}$$

As we can easily see from (2.7), mutual information is always non-negative:

$$I(X,Y) \geq 0. \tag{2.10}$$

In the extreme case of independence, mutual information is zero. The more dependent X and Y are, the larger the difference between the sum of their entropies and their joint entropy becomes. Thus, two random variables with high mutual information are more statistically dependent than two variables with lower mutual information.

Mutual information is also symmetric and exhibits some other interesting properties:

$$I(X,Y) = I(Y,X) \tag{2.11}$$

$$I(X,X) = h(X) \tag{2.12}$$

$$h(AX) = h(X) + \ln|\det(A)| \tag{2.13}$$

$$I(X,AY) = I(X,Y), \tag{2.14}$$

where A is a nonsingular square matrix.

2.2 Non-parametric Density Estimation

Frequently the density function of a random variable is unknown in its analytical form. It is then desirable to obtain an estimate of the density which fits the data

in a satisfactory way. Density approximation can be done in a variety of ways, which fall into two large categories, *parametric* and *non-parametric* estimates. A parametric estimate is essentially a family of allowable density functions with similar analytical form. For example, normal density functions with the mean and variance treated as parameters and polynomials of a certain type belong to this category. Although there is flexibility in the selection of the parameters, a parametric estimator is always restricted to only a family of functions and therefore cannot be good for the approximation of a wide range of diverse densities.

Consequently, when there is no a priori knowledge of the form of the density function, or when we want to use the same scheme for the approximation of a diverse set of densities, it is better to resort to non-parametric methods. Among the most widely used non-parametric estimators are the so-called *Parzen estimators*, or *kernel estimators*. The general expression for kernel estimators can be written as

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_N^d}K(\frac{x-x_i}{h_N}), \tag{2.15}$$

where $x_1, \ldots, x_N$ denote a random sample drawn from the distribution of random variable X, h_N is a scalar depending on N and d is the dimensionality.

An equivalent expression for the kernel estimator can be obtained by substituting $K_{h_N}(x) = K(x/h_N)/{h_N}^d$:

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}K_{h_N}(x-x_i). \tag{2.16}$$

K_{h_N} is called the *window function* or *kernel function*. One requirement for the window function is that it should be a legitimate density function, i.e. that

$$K_{h_N}(x) \geq 0 \tag{2.17}$$

and

$$\int K_{h_N}(x)\,dx = 1. \tag{2.18}$$

These combined with (2.16) ensure that the estimator is also a legitimate density function.

In the literature, there are various formulations and proofs of convergence – i.e. proving that as $N \to \infty$ the estimator approaches $p(x)$ on average (see [6], [10] etc.). The conditions for convergence also vary, but in practice the window functions that are mostly preferred have specific characteristics. They are usually selected to have a peak at the origin and to diminish as $|x|$ increases, so that the estimate has a higher value wherever there is higher concentration of samples. Furthermore, as the sample size increases the window should become narrower, but not too narrow. Among the window functions that adhere to the above, the *Gaussian* or *normal* distribution is one of the most popular.

The Gaussian distribution, centered at the origin, is defined by

$$G_\psi(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\psi)^{\frac{1}{2}}} e^{-\frac{1}{2}x^T \psi^{-1} x}, \tag{2.19}$$

where the parameter ψ is a $d \times d$ positive definite, symmetric matrix (called the *covariance matrix*). The covariance matrix determines the size and the orientation of the Gaussian. More specifically, when the covariance matrix is diagonal the Gaussian window is oriented symmetrically with respect to the coordinate axes and the diagonal elements indicate the breadth of the window along each of these axes. Decreasing one of the diagonal elements narrows the Gaussian along the corresponding dimension and heightens its peak (and conversely). Furthermore, non-diagonal covariance matrices correspond to rotations of these Gaussian windows in $\mathbb{R}^d$.

With a Gaussian as window function, the Parzen estimator now becomes

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} G_\psi(x - x_i). \tag{2.20}$$

It is clear that, depending on the covariance ψ, we can get an entire family of functions estimating the actual density. However, most of the estimates are not sufficiently good. If the elements of ψ are too large, then the Gaussians are too wide and the estimate shows too little variation to be useful. If, on the other hand, the variances are too small, then the Gaussians are too narrow and the estimate is noisy – it does not interpolate, it merely fits the sample. Therefore, attention should be paid to ensure that the covariance matrix belongs to some intermediate range between the two extremes.

2.3 Vision Background

We end the chapter by presenting some basic background knowledge from computer vision, necessary for the rest of the thesis.

Image Formation

One common way to model the formation of an image is to map the three-dimensional space of the real world to the image plane through *perspective projection* ([8]). That is, the projection of a visible point from a scene lies on the line connecting the point with the camera. Most of the time, it is convenient to choose a coordinate system with the origin at the camera, the xy-plane parallel to the image plane and the z-axis oriented towards the image plane. More formally, consider a point $x = (x_1, x_2, x_3)$ in space. Its projection $x' = (x'_1, x'_2)$ on the image is given by the equation

$$(x'_1, x'_2) = \frac{f}{x_3}(x_1, x_2), \tag{2.21}$$

where the constant f is the distance of the camera from the image plane (also called *focal length*).

Having established the correspondence between world points and image points, we can define an imaging function that associates each point x with the image intensity $v(x)$ at the projection x'. For a given image and an object at a given position, the

specific function v that could have produced the image from the object by perspective projection – ignoring for the moment the plausibility of such a function – is a known one. However, the problem of finding a global imaging function, i.e. of being able to produce the image of any object under any conditions, is practically impossible, the reason being that it is hard to assess the contribution of certain parameters influencing image formation. Furthermore, even if the generic function were known, estimating some of these parameters could be hard, if at all possible.

One important factor in image formation is the orientation of the surface at the point viewed, or equivalently the unit normal to the surface, $u(x)$. Intensity depends on the emittance angle, formed by the unit normal and the viewing direction, and on the angles formed by the unit normal and the incident light rays. A second parameter is the *albedo* $\rho(x)$, a factor indicating the fraction of incident irradiance reflected by the surface patch. The remaining parameters are external factors, with illumination being the most important. This term accounts for the way in which light falls on the surface, including direction of incident light rays, radiance of the light sources etc.

A simplifying assumption usually made is that the light sources and the camera are at a large distance from the object relative to its size. This allows the direction of each light source and the viewing direction to be considered constant over the points of the object. Thus, $u(x)$ alone can determine the emmitance and incident angles. From now on, this assumption will be made, since it allows us to concentrate on the contribution of $u(x)$ and $\rho(x)$. Still, the imaging function is too complicated to be generically determined, the difficulty lying in the remaining factors.

A special case yielding a simple imaging model is that of a single point source illuminating a *Lambertian surface.* A Lambertian surface reflects all incident light and can approximate well enough the reflectance properties of matte surfaces. It can be shown ([8]) that brightness is then proportional to the cosine of the incident angle (*Lambert's law*):

$$v(x) = |E\rho(x)(\hat{s} \cdot u(x))|. \tag{2.22}$$

Here, E is the radiance of the light source and $\hat{s}$ the unit vector in the direction of the source. Constant factors without any importance have been omitted from the equation. In the presence of many light sources the effects are simply superposed:

$$v(x) = |\sum_i E_i \rho(x)(\hat{s}_i \cdot u(x))|. \tag{2.23}$$

Since Lambert's law offers a simple but often realistic model, it will be used for constructing the synthetic experiments of Chapters 3 and 4.

Chapter 3

Using Mutual Information for Alignment

The concept and derivations of alignment by maximization of mutual information are largely based on the ideas and the discussion appearing in [11]. Viola uses Parzen estimators to obtain a generic estimate of the entropy of a distribution, the Empirical Entropy Manipulation and Analysis (EMMA) estimate, from a sample of the distribution. He then applies the idea of maximization of mutual information to cases of alignment, both alignment of image to image and alignment of 3D model to image. With this work as a starting point, this chapter describes the motivation behind mutual information, how it can be used for alignment and explains how it works.

3.1 Alignment

The central problem of this thesis is the alignment of three-dimensional objects to their images. When performing alignment, we are given a model of an object and an image taken after a rigid transformation was applied to the object. The objective is to determine this transformation by comparing the given image with the object model.

The type of object model we are going to use consists of points on the object's surface along with other information about the object, such as corresponding normals,

albedos etc. A good model should contain sufficiently many points, distributed as uniformly as possible over the object's surface, since it is desirable that there are no large areas uncovered by model points.

Knowledge of the model allows us to predict the position of the object under any rigid transformation (or *pose*) applied to the object. Such a transformation $T(x)$ can be expressed as the composition of a rotation and a translation:

$$T(x) = Mx + t. \tag{3.1}$$

Mx is the rotation part, which means that M is a 3×3 orthogonal matrix with determinant 1, and t is the 3×1 translation vector.

For most everyday objects, the human visual system has little difficulty in determining reliably and efficiently the aligning T, but the task proves to be really hard for computer programs. When shown two images of an object under different poses, people can roughly visualize the relation between the poses. Yet there is no thorough knowledge of the image formation process nor of the illuminating conditions. As a result, it is not possible to write a program that predicts what the image of the object would be at a certain pose, and of course neither the human brain does something similar. Thus, more indirect approaches should be sought.

As we saw in 2.3, image intensity should be a function of normals and albedos when other factors remain constant. Therefore, at the correct pose, intensity $v(T(x))$ on the given image is a function of the product of the transformed normals $Mu(x)$ and the albedos $\rho(x)$:

$$v(T(x)) = f(\rho(x)Mu(x)). \tag{3.2}$$

This implies that to solve the alignment problem it suffices to find an implicit way for comparing intensities on the image with normals and albedos on the object under the assumption of constant illuminating conditions. One such way could be to maximize a certain measure involving $v(T(x)), u(x)$ and $\rho(x)$ which we know that has a maximum at the correct pose. In the following sections it will be shown that mutual information

provides a good such measure.

3.2 The Connection of Mutual Information and Alignment

As we have seen in Section 2.1, the mutual information of two random variables describes how much statistical dependence exists between the variables. Specifically, high mutual information indicates that the variables are related in a simple way and low mutual information indicates that they are related in a more random way. Clearly, this intuition can be applied to alignment, whose solution is characterized by the existence of a strong functional relationship (Equation (3.2)). The case of high mutual information occurs at the aligning pose, where there is a strong functional relationship between the variables $v(T(x))$ and $\rho(x)u(x)$. At incorrect poses, on the other hand, the relationship breaks, becoming more and more "anarchic" as we move further away from the correct pose.

For example, in the ideal Lambertian case there is a linear relationship between $v(T(x))$ and $\rho(x)u(x)$. The alignment equation (3.2) in combination with (2.23) yields

$$v(T(x)) = |a^T \rho(x) u(x)|, \tag{3.3}$$

where a is the fixed vector such that $a^T = \sum_i E_i \hat{s}_i^T M$. Recall that $\hat{s}_i$ denotes the unit vector towards the direction of the i-th light source and E_i the radiance of the source. Note that Equation (3.3) holds only at the aligning T^*; at incorrect poses, there is no such linear relationship because the range of $v(T(x))$ and the domain of x may be different. The domain of x may change if the visible part of the model is affected by the transformation. Of course, there are objects which can exhibit a linearity like that of Equation (3.3) at more than one pose, even at an infinite number of poses (the rotation of a symmetric object with constant albedo would be a simple example). But then all these poses are visually plausible as aligning poses and can be accepted as satisfactory solutions in the absence of any additional knowledge.

Even when $v(T(x))$ and $\rho(x)u(x)$ are related through a different type of function than that of (3.3), the relationship should be stronger at the aligning transformation. Moving away from the aligning transformation, there are more and more discrepancies, since the correspondence between intensities and model properties is shifted, and the functional pattern is gradually dispersed. Consequently, one should expect that the mutual information of $v(T(x))$ and $\rho(x)u(x)$ is maximized at T^*.

Maximization of mutual information has been frequently used for similar problems in recent years. It has been applied successfully to the problem of source separation, where the input and the output are related through a nonsingular square matrix A: $y = Ax$. As Bell demonstrated, maximizing $I(Y, X)$ can be used for recovering the input x from y ([2]). The situation in the problem of alignment, and particularly in the linear special case (3.3), is similar in that we want to increase the statistical dependence between input and output in the equation $v(T(x)) = f(\rho(x)u(x))$. This provides motivation for applying information maximization to alignment. However, alignment cannot be based on the source separation algorithm since here the mixing function f is not invertible.

In order to be able to make use of a probabilistic concept like mutual information, our random variables must first be defined. Thus, X will be considered a continuous random variable with uniform density, whose sample space is the surface of the three-dimensional model. The mutual information of $v(T(X))$ and $\rho(X)u(X)$ is simply

$$I(v(T(X)), \rho(X)u(X)) = h(v(T(X))) + h(\rho(X)u(X)) - h(v(T(X)), \rho(X)u(X)),$$

where $v(T(X))$ is only defined at the visible part of the transformed model. The differential entropy terms expand as follows (see 2.1):

$$h(v(T(X))) = -\int p(v(T(x))) \ln p(v(T(x)))\, dv$$

$$h(\rho(X)u(X)) = -\int p(\rho(x)u(x)) \ln p(\rho(x)u(x))\, d(\rho u)$$

$$h(v(T(X)), \rho(X)u(X)) = -\int p(v(T(x)), \rho(x)u(x)) \ln p(v(T(x)), \rho(x)u(x))\, dv\, d(\rho u).$$

Among these entropy terms, $h(\rho(X)u(X))$ is known to be constant, because the model (at the original position) is fixed. On the contrary, $h(v(T(X)))$ and $h(v(T(X)), \rho(X)u(X))$ may vary, because when the transformation changes, the visible part of the transformed model and its projection to the image may change. Thus, for information maximization it suffices to maximize

$$h(v(T(X))) - h(v(T(X)), \rho(X)u(X)). \tag{3.4}$$

Except for the presence of albedos, this is the formulation of information maximization found in [11]. It is natural to incorporate albedo as a coefficient of $u(x)$ although other ways would probably work as well. For example, the mutual information $I(v(T(X))/\rho(X), u(X))$ would be appropriate for alignment, too. On the other hand, $I(v(T(X)), z(X))$, with $z(X) = \begin{pmatrix} \rho(X) \\ u(X) \end{pmatrix}$ being the joint vector of albedo and normal, is a bad choice, since the input variables are then intermingling with each other – mutual information cannot handle such problems.

3.3 Estimating Entropy

We now present the derivation of the EMMA[1] estimate of mutual information proposed in [11].

3.3.1 General Entropy Estimation

In order to be able to estimate mutual information, it is necessary to have estimates of the distributions of the random variables involved. Supposing a random sample A of cardinality N_A has been drawn from random variable X, the method of Parzen estimators can be used to approximate distributions and entropies. Thus, for a random

[1] Viola calls this method of approximating entropies "EMpirical entropy Manipulation and Analysis", in abbreviation EMMA.

variable $g(X)$, the estimate obtained is

$$\hat{p}(g(x)) = \frac{1}{N_A} \sum_{x_j \in A} G_\psi(g(x) - g(x_j)).$$

Now, entropy is an expectation (Eq. (2.6)), so it can be approximated by a sample mean taken over a second sample B of cardinality N_B:

$$h(g(X)) = E[-\ln p(g(X))] \simeq -\frac{1}{N_B} \sum_{x_i \in B} \ln p(g(x_i)).$$

Replacing p with its estimate $\hat{p}$, we get

$$h(g(X)) \simeq h^*(g(X)) = -\frac{1}{N_B} \sum_{x_i \in B} \ln[\frac{1}{N_A} \sum_{x_j \in A} G_\psi(g(x_i) - g(x_j))]. \tag{3.5}$$

What we will need is the derivative of this estimate with respect to some parameter r. To search a space of parameters for the values maximizing a certain mutual information measure, some gradient ascent method will be necessary, which means updating the parameters in proportion to the estimated gradient of mutual information.

From (3.5) we get

$$\frac{\partial h(g(X))}{\partial r} \simeq \frac{\partial h^*(g(X))}{\partial r} = -\frac{1}{N_B} \sum_{x_i \in B} \frac{\sum_{x_j \in A} \frac{\partial}{\partial r} G_\psi(g(x_i) - g(x_j))}{\sum_{x_j \in A} G_\psi(g(x_i) - g(x_j))}.$$

Since ψ is a symmetric matrix,

$$\frac{\partial}{\partial r} G_\psi(y) = -G_\psi(y) y^T \psi^{-1} \frac{\partial y}{\partial r}$$

and therefore,

$$\frac{\partial h^*(g(X))}{\partial r} = \frac{1}{N_B} \sum_{x_i \in B} \frac{\sum_{x_j \in A} G_\psi(g(x_i) - g(x_j))[g(x_i) - g(x_j)]^T \psi^{-1} [\frac{\partial g(x_i)}{\partial r} - \frac{\partial g(x_j)}{\partial r}]}{\sum_{x_j \in A} G_\psi(g(x_i) - g(x_j))}$$

or

$$\frac{\partial h^*(g(X))}{\partial r} = \frac{1}{N_B} \sum_{x_i \in B} \sum_{x_j \in A} W(g(x_i), g(x_j))[g(x_i) - g(x_j)]^T \psi^{-1} \cdot [\frac{\partial g(x_i)}{\partial r} - \frac{\partial g(x_j)}{\partial r}], \tag{3.6}$$

where

$$W(g(x_i), g(x_j)) = \frac{G_\psi(g(x_i) - g(x_j))}{\sum_{x_k \in A} G_\psi(g(x_i) - g(x_k))}.$$

Observe that $W(g(x_i), g(x_j)) > 0$ and that, in general, $W(g(x_i), g(x_j)) \neq W(g(x_j), g(x_i))$. Intuitively, $W(g(x_i), g(x_j))$ measures the relative proximity of $g(x_j)$ to $g(x_i)$. It has a value close to one when $g(x_j)$ is much closer to $g(x_i)$ than any of the other sample points from $g(A)$. Conversely, it has a value close to zero when other sample points are much closer to $g(x_i)$. The rest of the terms in (3.6) comprise the derivative of a symmetric product that provides a measure of squared distance between the sample points:

$$[g(x_i) - g(x_j)]^T \psi^{-1} [\frac{\partial g(x_i)}{\partial r} - \frac{\partial g(x_j)}{\partial r}] = \frac{\partial}{\partial r} \{\frac{1}{2} [g(x_i) - g(x_j)]^T \psi^{-1} [g(x_i) - g(x_j)]\}.$$

Therefore, $h^*(g(X))$ can be reduced (or increased) by reducing (increasing) the distance between neighboring points – pairs having $W \simeq 1$. This conforms with the intuition that shrinking clusters of neighbors reduces entropy whereas expanding them increases entropy.

3.3.2 Derivation for the Case of Alignment

Equation (3.6) can provide a way to estimate the derivative of mutual information of model and image, as defined in (3.4). This derivative is needed for the gradient ascent in the aligning algorithm. In this context, the parameter r represents one of the translation or rotation parameters of the transformation T.

First, let us make the following definitions:

$$n_i = \rho(x_i)u(x_i)$$

$$v_i = v(T(x_i))$$

$$w_i = \begin{pmatrix} v_i \\ n_i \end{pmatrix}$$

Denote also the covariance matrices of the Gaussian kernels used in the estimation of $h(v(T(X)))$, $h(v(T(X)), \rho(X)u(X))$ by ψ_v, ψ_{vu} respectively. Using the same samples A, B for both entropy terms, we obtain

$$\begin{aligned}\frac{\partial}{\partial r} I(v(T(X)), \rho(X)u(X)) \simeq & \frac{1}{N_B} \sum_{x_i \in B} \sum_{x_j \in A} W_v(v_i, v_j)(v_i - v_j)\psi_v^{-1}(\frac{\partial v_i}{\partial r} - \frac{\partial v_j}{\partial r}) \\ & -\frac{1}{N_B} \sum_{x_i \in B} \sum_{x_j \in A} W_{vu}(w_i, w_j)(w_i - w_j)^T \psi_{vu}^{-1}(\frac{\partial w_i}{\partial r} - \frac{\partial w_j}{\partial r}),\end{aligned} \tag{3.7}$$

where

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

$$W_{vu}(w_i, w_j) = \frac{G_{\psi_{vu}}(w_i - w_j)}{\sum_{x_k \in A} G_{\psi_{vu}}(w_i - w_k)}.$$

Note that the variables $\rho(x), u(x)$ are the properties of the model at its original position and therefore are independent of the transformation parameter r. Consequently, $\frac{\partial n_i}{\partial r} = 0$ and (3.7) becomes

$$\begin{aligned}\frac{\partial}{\partial r} I(v(T(X)), \rho(X)u(X)) \simeq & \frac{1}{N_B} \sum_{x_i \in B} \sum_{x_j \in A} (\frac{\partial v_i}{\partial r} - \frac{\partial v_j}{\partial r})[W_v(v_i, v_j)(v_i - v_j)\psi_v^{-1} \\ & -W_{vu}(w_i, w_j)(w_i - w_j)^T c_{vu}],\end{aligned} \tag{3.8}$$

where c_{vu} denotes the first column of ψ_{vu}^{-1}.

The above equation involves the gradient of intensity with respect to the trans-

formation parameters. This gradient is given by the following formula:

$$\frac{\partial v_i}{\partial r} = [\nabla v(T(x_i))]^T \frac{\partial}{\partial r} T(x_i). \tag{3.9}$$

The terms $T(x_i)$ and $\frac{\partial}{\partial r}T(x_i)$ can be accurately computed from the transformation parameters. If the image gradient is known, then $\nabla v(T(x_i))$ can be computed too from the projection equation (2.21). In the next section, we proceed to examine how well the above formulation performs, by testing it on a few simple objects.

3.4 Application to Geometrical Objects with Albedos

3.4.1 Construction of Synthetic Test Cases

To investigate how information maximization can be applied to alignment of three-dimensional models to images, we have developed models of simple geometrical objects, cubes and spheres that can be generated by the program during execution.

A model of an object was constructed by selecting a set of points from the theoretical surface in a way as uniform as possible (and non-random). The model consists of this set of points, along with the associated normals and albedos at the points. With a sufficiently high number of points (usually around 300,000), a good model can be obtained.

The images used in the experiments are grayscale images whose brightness varies in the range of integers from 0 to 255 and are produced synthetically from the models under the assumption of Lambertian surfaces. What we want is an information maximizing algorithm which, given the image produced from a model under a transformation T^*, starts from an arbitrary T and updates it until it converges to T^*.

3.4.2 Information Maximization Using the Image Gradient

Before putting Equation (3.8) to the task of alignment, several details need to be clarified. First of all, it is necessary to have a good estimate of the gradient of intensity with respect to the image coordinates, since it enters the computation of $\nabla v(T(x_i))$ in (3.9). A further relevant observation to be made is that any method using Equation (3.8) for information maximization should rely largely on pairs of points whose intensity gradients differ significantly. In contrast, pairs of points with similar intensity gradients have a much smaller contribution. In particular, those image points where an abrupt change of intensity occurs (usually at discontinuities of albedos or normals) have a large contribution, because their image gradient surpasses a lot that of points in smooth regions of the image. Thus, Equation (3.8) exploits any useful information that might be hidden in such discontinuous "edges". On the other hand, this implies higher sensitivity to noise contained in the image gradient and to the quality of the estimation method used.

What makes gradient estimation a subtle task is the fact that the image is discretized and, consequently, an interpolation scheme is required. We have chosen the same interpolation method both for making variable v continuous and for estimation of its gradient. Among the methods we have considered, the one that seems to give the most satisfactory results is also the most elaborate one and depends on the exact position of the point on the image. The interpolated value is a linear combination of the intensity values at the pixel containing the point, the closest neighboring pixel and the closest diagonal pixel. The advantage of this technique is the incorporation of intensity change in diagonal directions, apart from the horizontal or vertical directions, a feature that most of the time enhances gradient ascent.

The second issue of concern is the representation of the transformation T. In fact, there can be as few as 7 transformation parameters if the rotation part is expressed as a *quaternion*. A quaternion is simply a vector of 4 real numbers and can represent a three-dimensional rotation in the following way: a rotation of angle θ around a unit vector $(v_1, v_2, v_3)^T$ corresponds to the quaternion $Q = (\cos\frac{\theta}{2}, \sin\frac{\theta}{2}v_1, \sin\frac{\theta}{2}v_2, \sin\frac{\theta}{2}v_3)$.

Furthermore, such a quaternion is normalized, i.e. has length equal to one. The rotated image of a point can then be computed from the parameters of Q by a slightly complex formula (see [8], pp. 437-438). Normalized quaternions appear in all of the experiments that will be presented, since their concise representation for rotations is computationally efficient and reduces the dimensionality of the parameter space. Translations are represented as triple vectors in the common way.

Finally, some comments should be made on the process of selecting samples A and B of Equation (3.8). To reduce the computational overhead of selecting the samples, a strategy similar to *cross-validation* can be used. Instead of selecting two different samples ($N_A + N_B$ points per iteration), sample B alone can be used for both Parzen estimation and the sample mean of $\ln p$ (Equation (3.5)). Thus, for each i, sample A is taken equal to $B - \{x_i\}$, so that the estimator for $p(g(x_i))$ is still unbiased. An interesting note here is that even if A, B were independent and the Parzen estimate unbiased, the EMMA estimate would not be unbiased. This is due to the logarithm in the expression (3.5) and can be easily proved (see [11], p. 63). As we shall see in the next sections, these imperfections do not affect the usefulness of EMMA for information maximization.

3.4.3 Experiments with a Cube

The first example we will focus on is the case of a cube with constant albedo on each face (see Figure 3-1). The image of the cube was produced under ideal Lambertian conditions, so that image intensity is constant inside each of the three regions corresponding to the faces of the cube. In some sense, the goal is to classify points with the same normal in the same region of constant intensity. Consequently, albedo is not really important in this case, because only the intensity value at each region depends on it.

Figures 3-2, 3-3 depict the joint distribution of model and image at two different poses, using the same sample points x_i. At the correct pose, points w_i from the joint distribution appear at only three locations, whereas at an incorrect pose there are misclassifications and the number of locations increases. Since the relationship

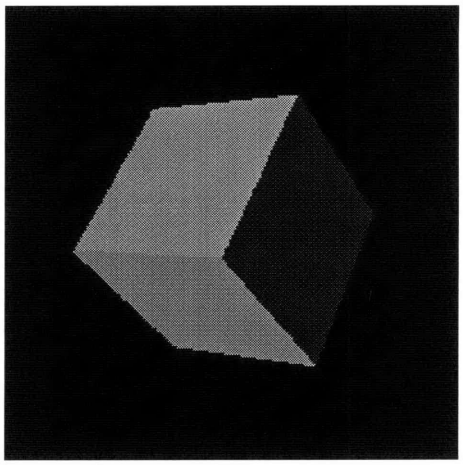

Figure 3-1: Image of the cube.

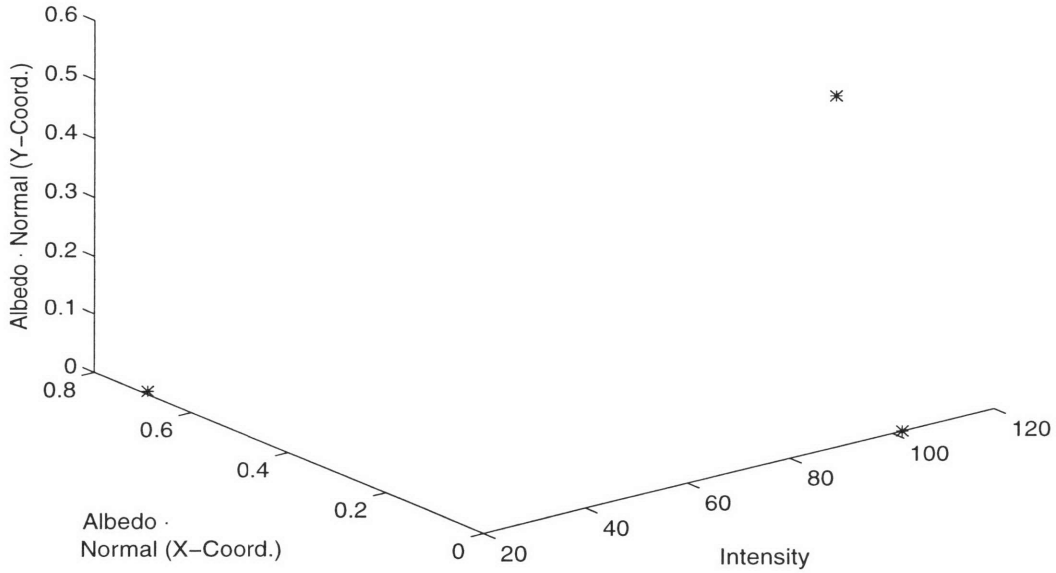

Figure 3-2: Plot of $v(T(x)), \rho(x)u_1(x)$ and $\rho(x)u_2(x)$ at the correct pose. Since only half the cube is visible, the z-coordinate of the normal is uniquely determined by the other two coordinates.

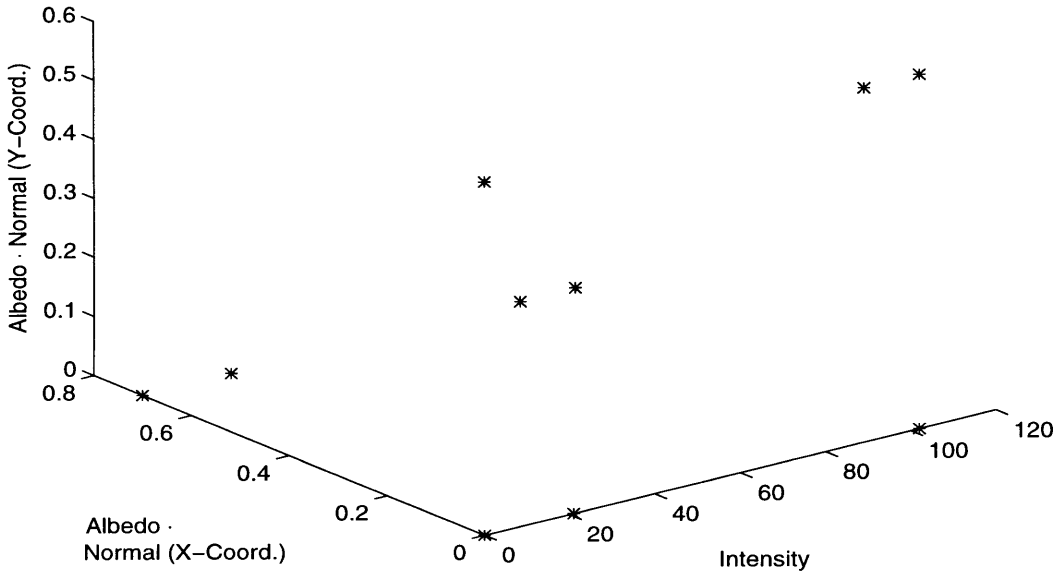

Figure 3-3: Plot of $v(T(x))$, $\rho(x)u_1(x)$ and $\rho(x)u_2(x)$ at an incorrect pose. Since only half the cube is visible, the z-coordinate of the normal is uniquely determined by the other two coordinates.

is simplest at the correct pose, joint entropy should be lowest there, which is an indication that mutual information is maximum at that point. With appropriate variances, the EMMA estimate should have a maximum there as well.

Figure 3-4 shows a graph of the mutual information estimate versus the first parameter of the rotation quaternion (the other parameters remaining fixed). A fixed sample of 100 points was used and the variances were chosen to be

$$\psi_v = 900,\ \psi_{vu} = \begin{pmatrix} 500 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 \end{pmatrix}.$$

Since the joint distribution consists of a few points only, any orientation is meaningless

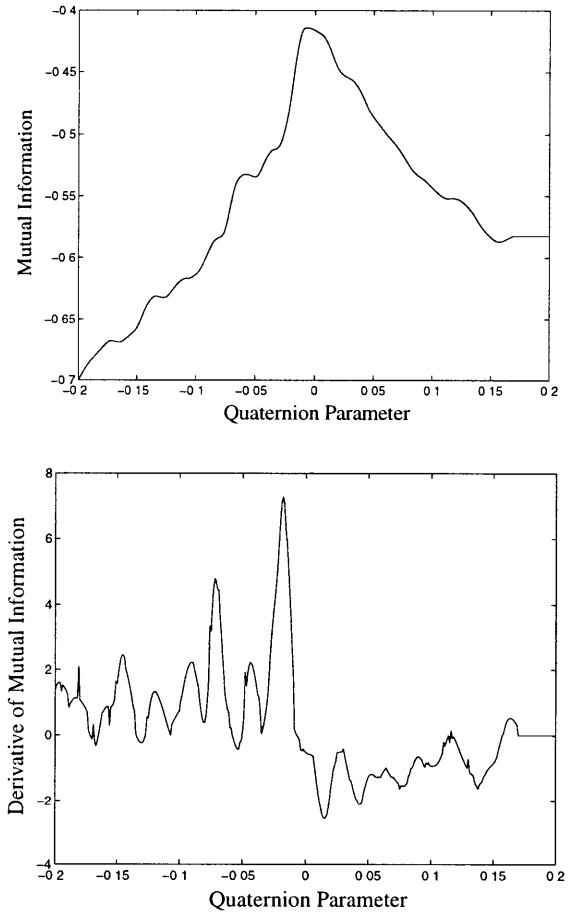

Figure 3-4: On top is a plot of mutual information versus the first quaternion parameter. The other parameters remain fixed at the values of the aligning quaternion. Position 0 on the x-axis corresponds to the aligning pose. The y-axis shows mutual information minus the (constant) entropy of the model. On bottom is a plot of the partial derivative of mutual information with respect to the first quaternion parameter.

and ψ_{vu} can simply be a diagonal matrix.

The plot indicates that there is a strong maximum very close to the aligning pose and, furthermore, that the terrain is not characterized by dangerous local maxima. In addition, the partial derivative of mutual information with respect to the quaternion parameter is predominantly positive before and predominantly negative after the aligning pose. Equation (3.8) describes how this works. The term $(\frac{\partial v_i}{\partial r} - \frac{\partial v_j}{\partial r})W_v(v_i, v_j)(v_i - v_j)\psi_v^{-1}$ tends to increase the distance between intensity values when these are similar. In this case it has the effect of moving the points towards one of the boundaries, regardless of whether they are correctly classified on the same face or not. The second term $-(\frac{\partial v_i}{\partial r} - \frac{\partial v_j}{\partial r})W_{vu}(w_i, w_j)(w_i - w_j)^T c_{vu}$ is equal to $-(\frac{\partial v_i}{\partial r} - \frac{\partial v_j}{\partial r})W_{vu}(w_i, w_j)(v_i - v_j)\psi_{vv}^{-1}$, where $\psi_{vv} = (\psi_{vu})_{11}$. It tends to equalize intensities of correctly classified points, i.e. to move points from the boundary to the interior of the face. Thus, a well-balanced choice of ψ_v, ψ_{vv} enables correctly classified points to stay together and misclassified points to move to another face of the cube. Recall also that the main contribution comes from image points near the edges. In fact, the contribution is zero at all other points (i.e. in the interior of the faces) since the image gradient is zero.

3.4.4 Experiments with a Sphere

Now, let us turn our attention to the case of a sphere with albedos. Obviously, alignment of a sphere would be meaningless without albedos, since it is a completely symmetric object. If albedo is constant all over the surface of the sphere, any pose is plausible – and indeed any pose satisfies some linearity of the form (3.3). Unlike the cube, alignment should largely rely on albedos and thus this is a good test case for their incorporation in the EMMA framework.

For the experiments, the surface of the sphere was divided in four equal regions of constant albedo in a tetrahedral configuration (see Figure 3-5). Note that the image of the sphere exhibits more variation than that of a cube, because of its curved shape. As a result, unlike the cube the joint distribution of model and image does not consist of just a few points, but of many scattered points, although certain patterns exist

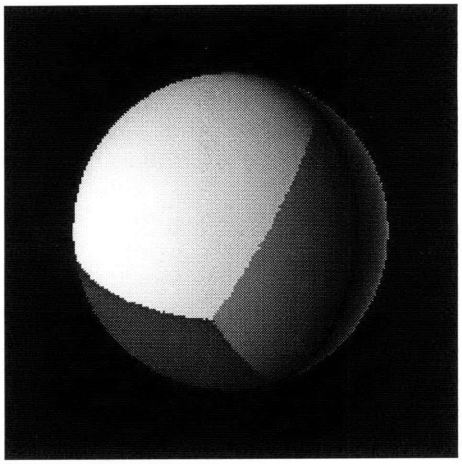

Figure 3-5: Image of the sphere.

here too. Figure 3-6 visualizes how the joint distribution becomes simplified at the aligning pose. This is just a projection of the four-dimensional space, but should be indicative because we have limited the range of the y-coordinate of the model variable by considering only values around 0.2.

Another way to look at the distribution is to renormalize it by an appropriate matrix multiplication of the model variable. As Equation (3.3) implies, it is possible with such an operation to make one coordinate equal to $v(T(x))$ (in absolute value) without changing the other two. Joint entropy then changes just by a constant (see Equation (2.13)), and thus two different poses can still be compared after multiplication with the same matrix. As expected, Figure 3-7 shows a simpler graph at the aligning pose and a more scattered one elsewhere. The same observation could be made even if the image were not produced in an ideal Lambertian way (Figure 3-8).[2]

The more complicated distribution in the case of a sphere also affects the selection of the covariance matrix ψ_{vu}. The Gaussian kernels now have to be oriented appropriately in the joint space, otherwise it is harder, if possible, to achieve satisfactory density estimation. Thus, a non-diagonal covariance matrix is preferable. For

[2]All points in the distribution were considered for these graphs – that is, no restriction was imposed to the coordinates.

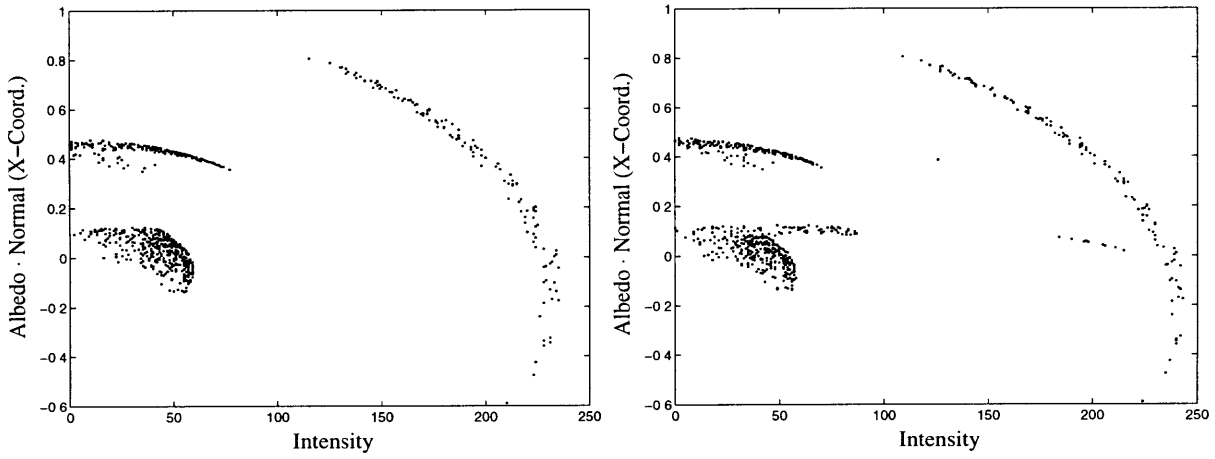

Figure 3-6: Scatter plots of intensity versus the x-coordinate of the model variable. On the left the aligning pose and on the right a non-aligning pose.

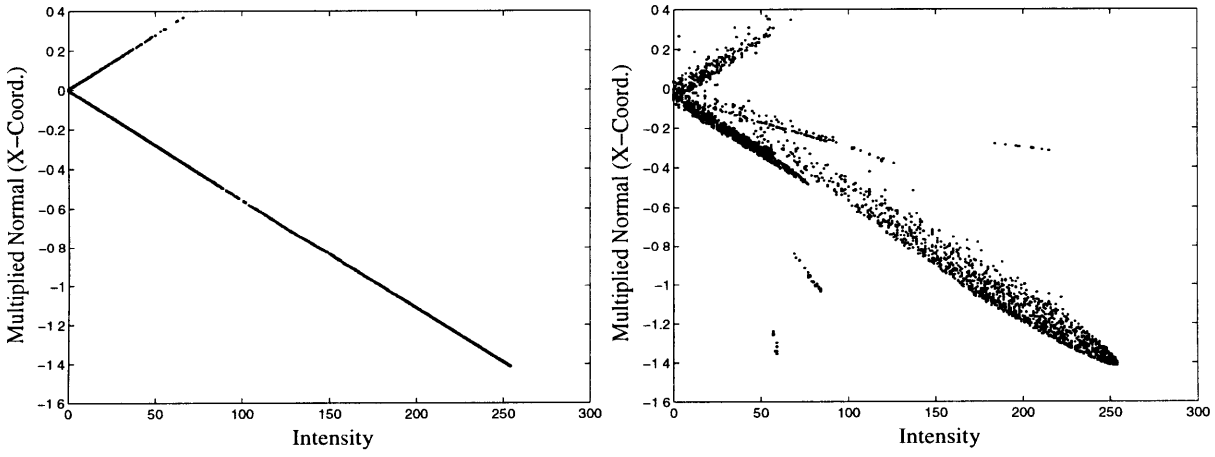

Figure 3-7: Scatter plots of intensity versus the normalized x-coordinate of the model variable. The model variable was multiplied by a matrix incorporating the aligning rotation and the source vectors. On the left is the aligning pose and on the right a non-aligning pose.

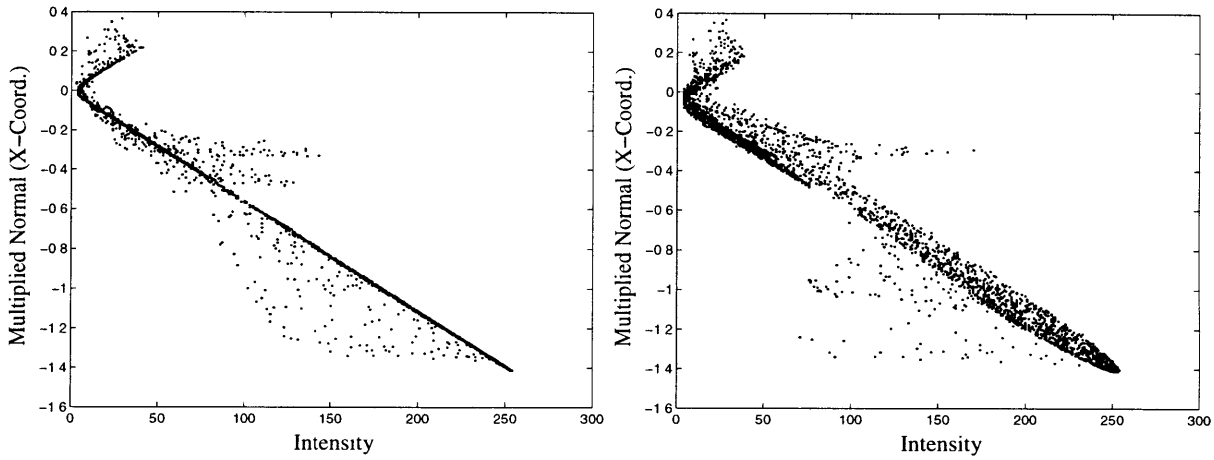

Figure 3-8: Scatter plots of intensity versus the normalized x-coordinate of the model variable. A blurred image of the model was used. On the left is the aligning pose and on the right a non-aligning pose.

instance, with a fixed sample of 100 points we can select variances

$$\psi_v = 3600,\ \psi_{vu} = \begin{pmatrix} 20500 & 20 & 0 & 0 \\ 20 & 15 & 1 & -1 \\ 0 & 1 & 0.2 & 0 \\ 0 & -1 & 0 & 0.2 \end{pmatrix}.$$

These values give good results for the specific sample we used and may be inappropriate for other samples, but still they can give an idea about the behavior of EMMA. The graph of the mutual information estimate versus the first quaternion parameter is shown in Figure 3-9.

As can be seen from the graph, there is a global maximum in the close vicinity of the aligning pose without local maxima posing any threats. This is confirmed by the partial derivative, which is predominantly positive before and predominantly negative after the aligning pose. Compared to the case of the cube, the surface seems to be slightly less smooth and the derivative seems to fluctuate more. Most likely this is due to the curved surface which, unlike the flat faces of a cube, ensures that all points contribute to the image gradient. The analysis of the behavior of the estimate has not changed significantly, however. In this case too, the gradient

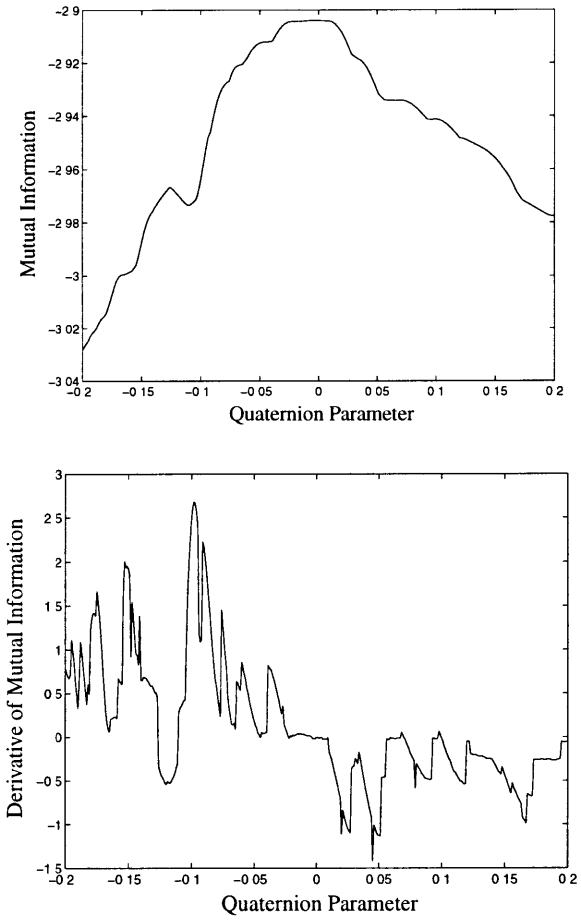

Figure 3-9: On top is a plot of mutual information versus the first quaternion parameter. The other parameters remain fixed at the values of the aligning quaternion. Position 0 on the x-axis corresponds to the aligning pose. The y-axis shows mutual information minus the (constant) entropy of the model. On bottom is a plot of the partial derivative of mutual information with respect to the first quaternion parameter.

$(\frac{\partial v_i}{\partial r} - \frac{\partial v_j}{\partial r})W_v(v_i, v_j)(v_i - v_j)\psi_v^{-1}$ reduces association in intensity value so that there can be classification changes, while the gradient $-(\frac{\partial v_i}{\partial r} - \frac{\partial v_j}{\partial r})W_{vu}(w_i, w_j)(w_i - w_j)^T c_{vu}$ tends to preserve and strengthen correct classification of points.

3.5 Stochastic Gradient

There remains now one last missing piece for a complete information maximization alignment algorithm. As it has been mentioned, the goal is some gradient ascent procedure based on the estimate (3.8) of the derivative of mutual information. That is, at the end of each iteration the transformation parameters are updated by adding the gradient of mutual information multiplied by a scalar called the *learning rate.* This leads to the update rule for each iteration:

$$r \leftarrow r + \lambda \frac{\partial I^*}{\partial r}, \tag{3.10}$$

where λ is the learning rate, r a transformation parameter and the sample B changes at every iteration. The important detail in this scheme is that $\frac{\partial I^*}{\partial r}$ is just a noisy estimate of the gradient of mutual information, not the actual value, which is not available anyway. The experiments of Sections 3.4.3 and 3.4.4 indicate that this estimate is a good one as far as maximization is concerned. But it is not clear whether such a stochastic gradient ascent should actually converge and under which conditions it does.

Stochastic gradient approximation has appeared in a diverse range of problems (see [7], [12], for example) and there have been theorems of convergence in the literature. In [11], stochastic convergence for the case of alignment is discussed and it is suggested that the stochastic gradient should work, even though some of the theoretical conditions may not strictly hold. A slight complication might be caused by one condition stating that the gradient estimate should be unbiased. This is not true, as we have already mentioned in Section 3.4.2, but it is possible to modify the definition of the entropy estimate in order to obtain an unbiased estimator (see [11], pp. 63-65).

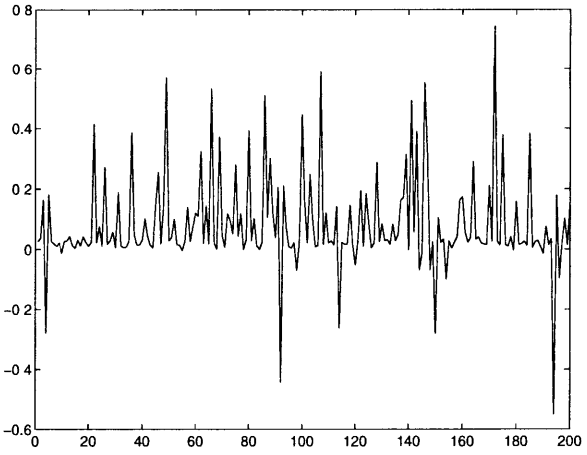

Figure 3-10: The estimate of the derivative of mutual information at 200 different samples. The derivative is with respect to the first quaternion parameter and at a fixed pose. Although the estimate is very noisy, the mean equals 0.0806 and is clearly positive, as the actual derivative should be.

However, the experiments we have performed indicate that even this modification is not really necessary. Our observations show only negligible differences in values between EMMA and the modified estimator and no serious effects on convergence or its speed.

In any case, the choice of stochastic maximization is probably the best available tool; it is also dictated by efficiency concerns. It turns out that increasing the sample size in order to improve the accuracy of the estimate is not a good idea. We would probably need thousands of sample points for a sufficiently reliable estimate. This is not possible because the complexity of computing the EMMA derivative is quadratic in N_B, as can be seen from Equation (3.8). For decent performance on a workstation, the sample size should be limited to no more than 50 points. Such a small sample results in high fluctuation in the estimate of the gradient (see Figure 3-10), but the average effect is in the right direction. An additional advantage of the stochastic gradient is that, because of the noise in it, there is much less probability of entrapment in local maxima than there is with the traditional gradient method. For all of the above reasons, stochastic gradient is the best method to choose in the current context.

In the next chapter, which presents some alignment experiments, we will be able

to verify how well this method performs in practice.

Chapter 4

Experimental Results

In this chapter, the alignment experiments we have performed are presented. We have tested the alignment algorithm on the same synthetic objects used in Sections 3.4.3 and 3.4.4. Any remaining implementation details are also clarified and results are briefly discussed for each test case.

4.1 Overview of Implementation

First, let us summarize the aligning procedure, recapitulating all relevant information from the previous chapter. Starting from an initial transformation, at each iteration of the algorithm the visible part of the model is sampled and the corresponding normals, albedos and intensities are obtained. With this information an estimate of the mutual information gradient is computed from (3.8) and the transformation is updated according to (3.10). This process is repeated until the transformation parameters have stabilized.

Though the algorithm is fairly simple, there are several implementation details requiring attention – some of them have already been discussed in Sections 3.4.1, 3.4.2. But nothing has been said e.g. about determining the visibility of the model every time the transformation is updated. For this purpose, the method called *z-buffering* was used. With z-buffering, all points projecting on an image pixel are sorted by their z-coordinate and the one with the smaller value is selected as the

"visible point" for that pixel. The computational cost of this method would be disproportionate, however, since the model size is very large (in the order of 10^5). In fact, z-buffering can be performed very infrequently (every 250 iterations) without any real difference. The reason is that if the transformation does not change drastically between z-bufferings and the sample size is not large, then visibility can safely be considered almost fixed. Using z-buffering also implies that sampling is uniform with respect to the image coordinates. This contradicts the assumption of uniform sampling with respect to the model. However, which random variable, $v(T(X))$ or X, is considered uniform should not make a difference, as long as they are related through an invertible function.

Another issue is the choice of the learning rate. It is a good idea to use different learning rates for different transformation parameters instead of a unique rate. A change in the translation parameters has much less effect on the position of the model than an equal change in the quaternion parameters. Thus, we have to use larger learning rates for translation and smaller ones for rotation. In fact, the rate for z-axis translation was chosen to be much larger than the rate for the other translation parameters, because in all experiments the objects are located far from the image plane. It should also be noted that learning rates remained constant for a large number of iterations and then were significantly reduced, so that convergence closer to the maximum could be enabled.

In all cases a sample size of 30 points was chosen, so that the execution can be fast enough without compromising on correctness. Covariance matrices (ψ_{vu}, ψ_v) for each case were different, because the nature of the joint distribution differs from object to object. The sensitivity of the algorithm to ψ_{vu}, ψ_v varies from case to case as well. The values of the covariances were selected through trial and error, but there is a procedure which could be used for their estimation (see [11], pp. 67-68).

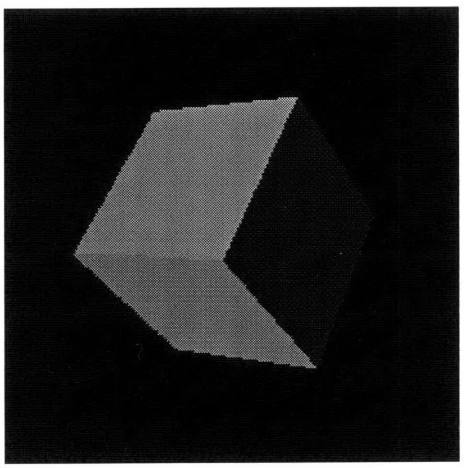

Figure 4-1: Target cube image.

4.2 Alignment of a Cube

The first object we used in our experiments is the cube of Section 3.4.3 (Figure 4-1). A rotation learning rate $\lambda_r = 0.0005$ and an xy-translation learning rate $\lambda_t = 5$ were used, while the z-translation learning rate, λ_z, was set equal to zero. It seems that during execution translation on the z-axis can have undesirable consequences and can lead away from the aligning pose. In this example, zooming out from the aligning pose should keep mutual information constant, since the probability distributions do not change, only the image area covered by the model. Thus, translation on the z-axis was ignored. [1] After the algorithm settled around the aligning pose, the learning rates were reduced by a factor of 5. Lastly, the variances selected were the following:

$$\psi_v = 10000,\ \psi_{vu} = \begin{pmatrix} 1000 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.4 \end{pmatrix}.$$

These provide only an example since the range of values that could be used with satisfactory results is fairly large.

[1] Such a problem could be avoided by an alternative formulation, like the one in [11], pp. 104-105.

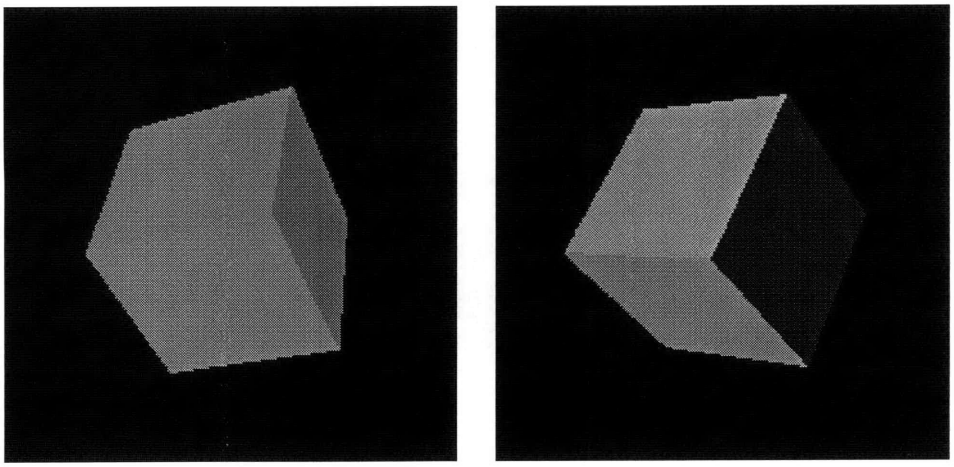

Figure 4-2: On the left an initial pose that differs from the correct pose in rotation only. On the right the final pose.

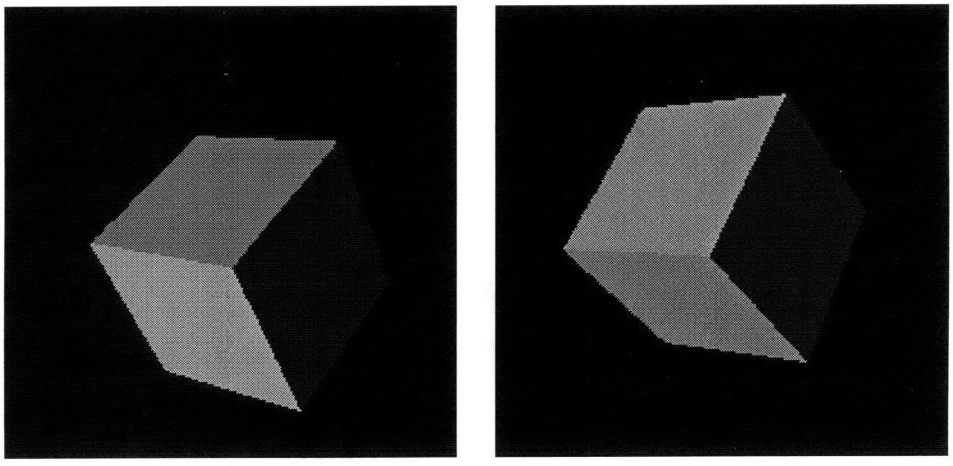

Figure 4-3: Initial and final pose. The initial pose differs from the correct pose in both rotation and translation.

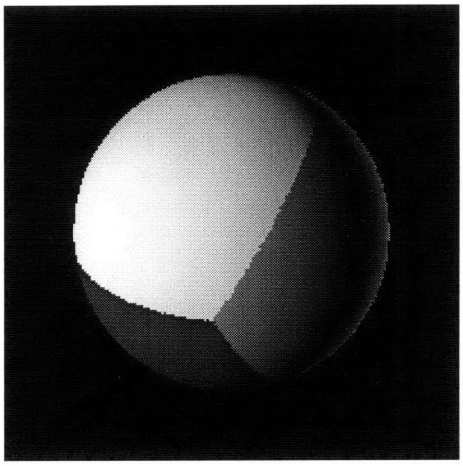

Figure 4-4: Target sphere image.

The experiments indicate that alignment is successful when starting inside a wide range around the aligning pose. Figure 4-2 shows an experiment from a rotated only pose, Figure 4-3 an experiment where rotation and translation are combined.[2] The approximate number of iterations needed to arrive in the vicinity of the aligning pose was 1000 in the first case and 2000 in the second. Trials with many initial transformations that are not too far from the correct pose have given similar results. Moreover, once the parameters reach near the correct pose, they remain there oscillating. In general, we have observed that the resulting translation is not as accurate as the resulting rotation and that the gradient with respect to translation is noisier, which makes convergence harder.

4.3 Alignment of a Sphere

The second set of experiments involves the sphere of Section 3.4.4 (Figure 4-4). The learning rates used were $\lambda_r = 0.001, \lambda_t = 20$ and $\lambda_z = 200$. We reduced these rates by a factor of 3 in the second stage of the gradient ascent. The covariance matrices

[2]To visualize the pose, the figures show synthetic images of the model.

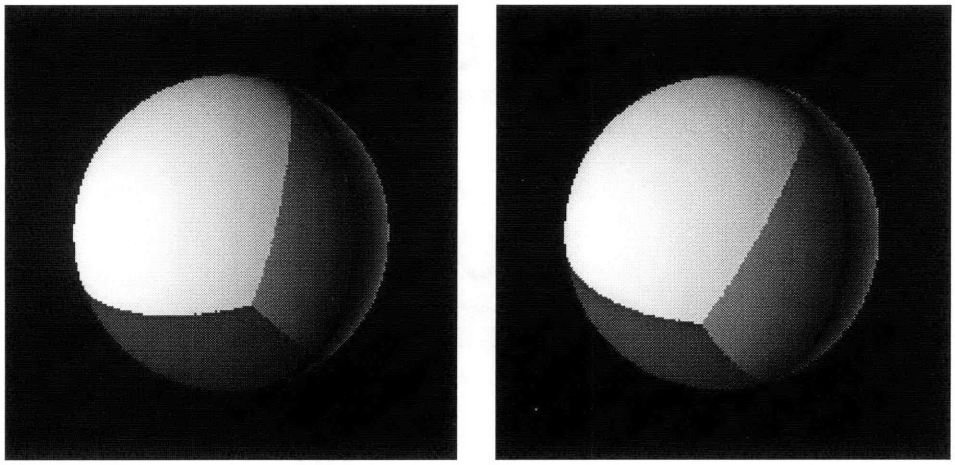

Figure 4-5: On the left an initial pose that differs from the correct pose in rotation only. On the right the final pose.

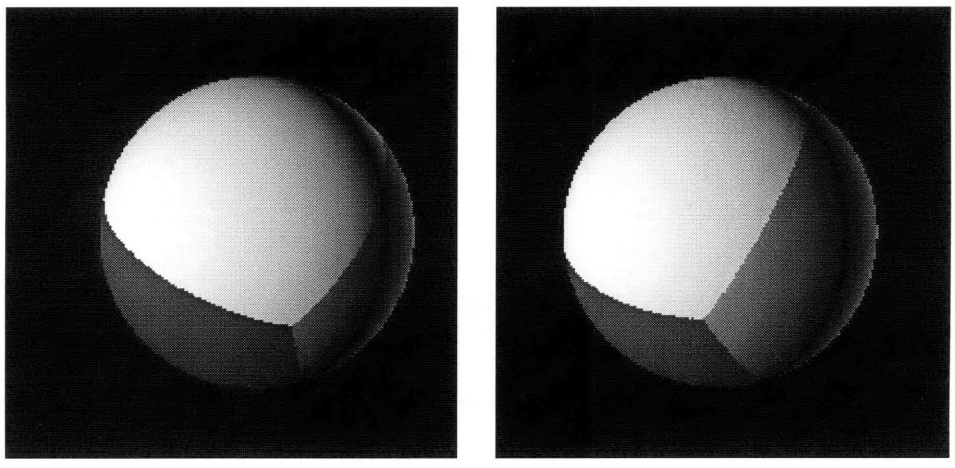

Figure 4-6: Initial and final pose. The initial pose differs from the correct pose in both rotation and translation.

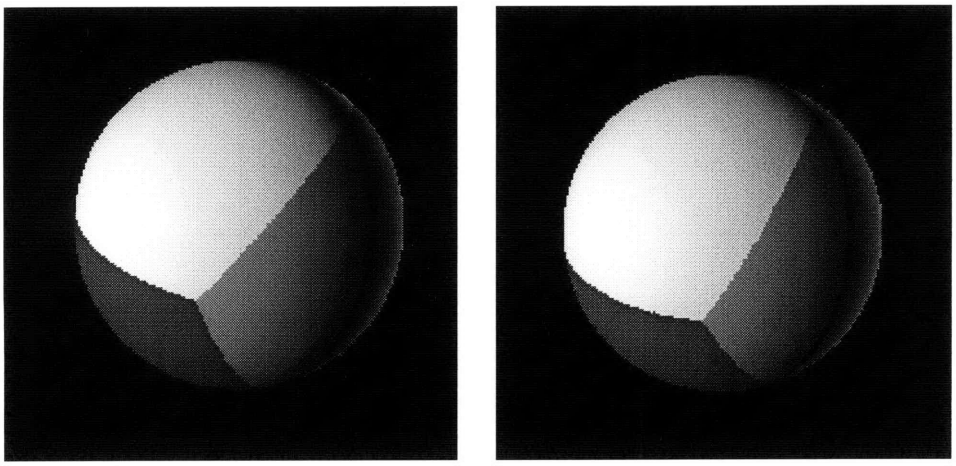

Figure 4-7: Initial and final pose. The model now starts from a position closer to the camera.

selected were

$$\psi_v = 250000,\ \psi_{vu} = \begin{pmatrix} 20100 & 65 & 3.5 & -4 \\ 65 & 75 & 5 & -7 \\ 3.5 & 5 & 1 & 0 \\ -4 & -7 & 0 & 1 \end{pmatrix}.$$

As before, this is only an example from an appropriate range of values, although this range is smaller than in the case of a cube. Furthermore, it is not clear whether we can simplify matters by selecting a diagonal ψ_{vu} – for example, the matrix obtained by removing the non-diagonal elements from the above ψ_{vu} does not align the sphere. Also note that the values for ψ_v and ψ_{vu} should be well balanced, so that the two entropy terms have balanced contributions to the estimation of mutual information.

Figures 4-5, 4-6, 4-7 demonstrate the results of the algorithm on three characteristic cases. The first and third cases needed about 3000 iterations until being attracted close to the maximum, while the second required about 2000 iterations. Repeated experiments with the sphere have shown convergence for a wide range of poses around the aligning pose. In this case too, the translation part is more problematic than the rotation part and the gradient with respect to translation behaves erratically.

Chapter 5

Conclusion

5.1 Thesis Summary

In this thesis, we have discussed an information maximization approach to a problem from computer vision, namely alignment of three-dimensional objects to images. We have used an already existing method of alignment by maximization of mutual information, which we have reformulated in order to be able to incorporate albedos. Furthermore, the theoretical behavior of the method was verified and interpreted in practice with the help of synthetic cubic and spherical objects. It has thus been demonstrated that the mutual information between model surface properties and image intensity has a maximum at the correct transformation. Finally, some experimental results of the execution of the stochastic gradient ascent algorithm maximizing mutual information were presented. In these, it is demonstrated that the algorithm performs alignment correctly.

5.2 Evaluation and Future Research

We have seen that there is strong motivation, at least in the Lambertian case, for applying information maximization to alignment. Moreover, these expectations have been verified in the experiments of Chapter 4. It seems that mutual information is very relevant to the human perception of alignment and of match between im-

ages, since it behaves in the same way. Above all, it does not require any special assumptions about the world or the lighting conditions, a fact that ensures success in a diverse set of situations. Thus, the same algorithm can be used for the alignment of different surfaces, both curved and polyhedral, without being influenced by the background, the illumination or the imaging technique. The reason why mutual information takes only the really relevant factors into account is that it can indicate functional dependence between random variables for unknown functions and that it makes minimal assumptions about the nature of these functions.

However, the experiments we have presented involve only synthetic objects and are based exclusively on the Lambertian model. This implies that more realistic situations need to be tested for a complete evaluation of the method. For example, we have confirmed with additional experiments that the algorithm also works when the image is filtered or corrupted with noise. Testing the method with real photographs of cubes and spheres would also be a further interesting task. As the alignment experiments included in [11] indicate, information maximization appears successful with many real life imaging situations. Nevertheless, this does not imply that mutual information can handle all possible problems. It provides a good measure for a specific category of functional relationships between signals, but it should not be used for other types of functions. Alignment, however, appears to be of a nature suitable for mutual information. Still, even in the case of alignment, there are counterexamples - e.g. the z-translation of the cube in Section 4.2.

Another important observation is that alignment is achieved only inside a range around the correct transformation. The outcome of some poses which the human eye can easily align is not correct, because of local maxima in the gradient ascent. This could be avoided perhaps by a scheme that starts from a variety of initial poses and selects the final pose with the highest mutual information estimate. Still another issue is that the algorithm is successful only with an appropriate choice of parameters. Reasonable learning rates have to be chosen, as in every gradient procedure. Furthermore, the covariance parameters have great influence on correctness and should belong to a certain range. The possible range of values as well as the sensitivity to

the parameters varies from case to case.

Finally, it should be noted that alignment appears in computer vision applications as a subproblem and is not sufficient in itself for undertaking visual tasks similar to those handled by the human visual system. Issues like obtaining a good model for the object to be aligned (especially when the object is not simple) or determining whether the object is present in the image should be addressed too. Thus, one direction for future work could be integration of the alignment algorithm with existing techniques for solving other related problems. On the other hand, alignment may become a source of inspiration for other vision problems, since research combining computer vision and information maximization is only a novel development. In general, information theory appears to be a powerful tool for tackling a diverse category of problems and provides useful intuition about some processes that are not well understood. Therefore, it is likely that further significant results of the cooperation between the two fields are going to appear in the future.

Bibliography

[1] S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, number 8. MIT press, 1996.

[2] A. Bell and T. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. Technical report, Computational Neurobiology Laboratory, The Salk Institute, 1995.

[3] A. Bell and T. Sejnowski. Edges are the 'independent components' of natural scenes. In *Advances in Neural Information Processing Systems*, number 9. MIT press, 1996.

[4] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley series in telecommunications, 1991.

[5] G. Deco and D. Obradovic. *An Information-Theoretic Approach to Neural Computing.* Springer-Verlag, 1996.

[6] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* Wiley, 1973.

[7] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Maxwell Macmillan International, 1994.

[8] B. Horn. *Robot Vision.* MIT Press, 1986.

[9] A. Papoulis. *Probablity, Random Variables and Stochastic Processes.* McGraw-Hill, 1991.

[10] R. Tapia and J. Thompson. *Nonparametric Function Estimation, Modeling and Simulation.* Society for Industrial and Applied Mathematics, 1990.

[11] P. Viola. *Alignment by Maximization of Mutual Information.* PhD thesis, Massachusetts Institute of Technology, 1995.

[12] M. T. Wasan. *Stochastic Approximation.* Cambridge University Press, 1969.