Geometry and Photometry in 3D Visual Recognition

by

Amnon Shashua

B.Sc., Tel-Aviv University, Israel (1986)
M.Sc., Weizmann Institute of Science, Israel (1989)

Submitted to the Department of Brain and Cognitive Sciences
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute Of Technology
November, 1992

Artificial Intelligence Laboratory, AI-TR 1401

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124. A. Shashua was also supported by NSF-IRI8900267.

Abstract

This thesis addresses the problem of visual recognition under two sources of variability: geometric and photometric. The geometric deals with the relation between 3D objects and their views under parallel, perspective, and central projection. The photometric deals with the relation between 3D matte objects and their images under changing illumination conditions. Taken together, an alignment-based method is presented for recognizing objects viewed from arbitrary viewing positions and illuminated by arbitrary settings of light sources.

In the first part of the thesis we show that a relative non-metric structure invariant that holds under both parallel and central projection models can be defined relative to four points in space and, moreover, can be uniquely recovered from two views regardless of whether one or the other was created by means of parallel or central projection. As a result, we propose a method that is useful for purposes of recognition (via alignment) and structure from motion, and that has the following properties: (i) the transition between projection models is natural and transparent, (ii) camera calibration is not required, and (iii) structure is defined relative to the object and does not involve the center of projection.

The second part of this thesis addresses the photometric aspect of recognition under changing illumination. First, we argue that image properties alone do not appear to be generally sufficient for dealing with the effects of changing illumination; we propose a model-based approach instead. Second, we observe that the process responsible for factoring out the illumination during the recognition process appears to require more than just contour information, but just slightly more. Taken together, we introduce a model-based alignment method that compensates for the effects of changing illumination by linearly combining model images of the object. The model images, each taken from a different illumination condition, can be converted onto novel images of the object regardless of whether the image is represented by grey-values, sign-bits, or other forms of reduced representations.

The third part of this thesis addresses the problem of achieving full correspondence between model views and puts together the geometric and photometric components into a single recognition system. The method for achieving correspondence is based on combining affine or projective geometry and optical flow techniques into a single working framework.

Thesis Advisor: Professor Shimon Ullman

Acknowledgments

First and foremost, thanks go to Shimon Ullman. I was fortunate to have Shimon as my advisor for the past six years, both at Weizmann and at MIT. Shimon allowed me to benefit from his knowledge, experience, financial support, and many of his original views while not constraining my creative efforts. I also gratefully acknowledge Shimon's influence on shaping my views in the fields of computer and human vision.

Thanks also to other members of the Brain and Cognitive Sciences department and the Artificial Intelligence Laboratory: to Tomaso Poggio for his generosity, continuous support and interest in my work, for countless illuminating discussions, and much more — but mostly for always treating me as a colleague and friend. To Whitman Richards for his never ending generosity and patience and for the many interesting "wednesday lunches" we had together. To Eric Grimson for the welg-group seminars and for patiently reading through my memos. To Berthold Horn for interesting discussions on issues related to photometry and geometry.

Thanks to the members of my thesis committee: to Tomaso Poggio who kindly agreed being the chairman, and to Ted Adelson, Ellen Hildreth, and Alan Yuille for making their time available to me. Special thanks to Tomaso Poggio and Ted Adelson for being my co-sponsors for the post-doctoral McDonnell-Pew fellowship.

Thanks to current and former members of the Center for Biological Information Processing: to Shimon Edelman who together with his wife Esti helped to make our first year in the U.S. very pleasant; to Daphna Weinshall for putting a kind word whenever possible and for being my Tennis partner; to Heinrich Bulthoff for many discussions on "Mooney" images, and to Norberto Grzywacz for always keeping spirits high.

For stimulating my interest in visual motion and structure from motion and for providing an exciting and inspiring summer, thanks to Peter Burt, Padmanabhan Anandan, Jim Bergen, Keith Hanna, Neil Okamoto and Rick Wildes at David Sarnoff Research Center.

Thanks to the vision group at IBM T.J. Watson Research Center: to Ruud Bolle, Andrea Califano and Rakesh Mohan for being generous enough to give me all the freedom in pursuing directions of research during my stay there at the summer of 1990; to Gabriel Taubin for many months of inspiring collaboration on combinatorial algorithms.

Many thanks to my fellow lab-mates at the Artificial Intelligence Laboratory: to Tanveer Syeda and Ivan Bachelder my office-mates, to David Jacobs for many inspiring chats on projective geometry and for his active role in the welg-group seminars, to David Clemens for making our life with Lisp Machines fairly enjoyable, to Anita Flynn for initiating many of the events that make this lab a fun place, to Pawan Sinha for always being helpful, to Brian Subirana for keeping spirits and "Crick" dinners up, to Davi Geiger for endless energy and enthusiasm; and to Ronen Basri, David Beymer, Thomas Breuel, Todd Cass, Ron Chaney, Frederico Girosi, John Harris, Ian Horswill, David Michael, Pam Lipson, Jose Robles, Kah Key Sung, Ali Taalebi, Sandy Wells, and Paul Viola who have made my stay here both enjoyable and inspiring.

Thanks to my fellow students at the department of Brain and Cognitive Sciences, especially to Sherif Botros, Diana Smetters, Eric Loeb and Randy Smith. Thanks to my fellow students at the Media Laboratory: Bill Freeman, Mathew Turk, Trevor Darrell and Eero Simoncelli.

Many thanks to the Artificial Intelligence Laboratory and to the department of Brain and Cognitive Sciences for providing a stimulating environment, especially to Janice Ellertsen who runs the BCS department, and to Liz Highleyman, Sally Richter and Jeanne Speckman at the AI lab.

To my parents, many thanks for their moral and financial support during what seems as the uncountable years of higher education.

Finally, to my wife without whom this thesis would not be possible, endless gratitude.

To Anat, with Love

Contents

Introduction

Chapter 1

The problem of visual object recognition is the focus of much interest in human and computer vision. The task seems very easy and natural for biological systems, yet has proven to be very difficult to place within a comprehensive analytic framework.

There are many aspects to the problem of recognition, many relevant sources of information, and apparently not a single widely accepted definition of what the problem is. For example, physical objects in the world can be identified based on various visual cues that include shape, color and texture. The images that an individual object can create depend on geometric properties, such as viewing position, on photometric properties such as the illumination conditions, and also on object characteristics such as the ability to change shape, having movable parts, and so forth. Objects often appear in the context of other visual information, such as when a scene contains multiple objects that are next to each other, or partially occluding each other. Objects can be classified as belonging to a general category or be identified as individuals. Finally, the kind of visual analysis that is employed in the process of recognition is not limited to the task of object recognition. Therefore, recognition may involve more than simply naming the object; it may also provide other information that is useful for motor interaction, following a path, and movements in the world in general.

The multitude of aspects to visual recognition and the considerable degree of abstraction associated with it implies that in order to make the problem amenable to analytic treatment, some form of problem simplification is required. In this thesis we are primarily concerned with shape-based recognition of individual three-dimensional (3D) objects from a single image of the object. The component within this context that we emphasize is that of dealing with the mathematical problem of understanding the relationship between objects in the world and their images. This component has two parts, geometric and photometric. The geometric part of the problem has to do with the relationship between different views of the same object produced by means of a central projection onto the image plane. The photometric part has to do with the relationship between images of the same object produced by changing the lighting conditions (level of illumination, positions and distributions of light sources).

We consider the case of recognition from full grey-level images and from reduced images (such as are produced by edge detection, or are binary images produced by threshold operation on the original image). Our definition of "success" is the ability to reproduce, or synthesize, a precise copy of the image in question from the model representation. This definition is adopted from the alignment approach to recognition.

1.1 Sources of Variability

One of the characteristic problems in visual recognition is the one-to-many mapping between an individual object in space and the images it can produce. As we move our eyes, change position relative to the object, or move the object relative to ourselves, the image of the object undergoes change. Some of these changes are intuitive and include displacement and/or rotation in the image plane, but in general the changes are far from obvious because of the nature of perspective projection from a 3D world onto a 2D plane. If the illumination conditions change, that is, the level of illumination, as well as the positions and distributions of light sources, then the image of the object changes as well. The light intensity distribution changes, and shadows and highlights may change their position. In general we may regard the one-to-many mappings as sources of variability that affect the kind of images that an individual object can produce. We distinguish four general sources of variability:

- Geometric: changes in the spatial location of image information as a result of a relative change of viewing position.
- Photometric: changes in the light intensity distribution as a result of changing the illumination conditions.
- Varying Context: objects rarely appear in isolation and a typical image contains multiple objects that are next to each other or partially occluding each other. Changes in the image can, therefore, occur by changing the context without applying any transformation to the object itself.
- Non-rigid Object Characteristics: these include objects changing shape (such as facial expressions), objects having movable parts (like scissors), and so forth.

The geometric source of variability has to do with the geometric relation between rigid objects and their perspective images produced under changing viewing positions (relative motion between the viewer and the object). This is probably the most emphasized source of variability and has received much attention both in the context of recognition and in the context of structure from motion. There are several approaches to this problem, depending on the model of projection that is assumed (orthographic or perspective), the object model representation (3D, or a number of 2D views), and the representation of structure (metric or non-metric). This is reviewed in more detail in Section 1.3, but we briefly mention here that in spite of extensive research in this area hard mathematical problems remain. For example, there is a lack of uniformity with respect to the model of projection, i.e., solutions are often approached by either assuming orthographic or perspective projection, but not both at the same time. Most of the research to date is focused on orthographic and parallel projections, where methods that assume perspective projection are often extremely sensitive to noise, require non-linear computations and do not fully address the issues in a comprehensive manner (i.e., necessity of calibration, the kind of metric or non-metric properties that are worth exploring, and so forth).

The photometric source of variability has to do with the relation between objects and the images they produce under changing conditions of illumination, i.e., changing the level of illumination, direction and number of light sources. This has the effect of changing the light intensity distribution in the image and the location of shadows and highlights. The dominant approach is to recover features from the image that are invariant to changes in illumination conditions. Under this approach the photometric source of variability turns into a question of image representation. The best known example of such features are step edges, namely, contours where the light intensity distribution changes abruptly from one level to another. Such edges are often associated with object boundaries, changes in surface orientation or material properties. The issue of image representation will be discussed in more detail in Chapter 4, but we can briefly mention here that the representation of edges and the invariance they provide is mostly sufficient for simple objects, such as polyhedrons and simple machine parts. Problems with the sufficiency of edge representation and its invariance against changing illumination arise with more complex objects, such as a face, a shoe, and so forth. In this case, we argue that a similar approach to that taken with the geometric source of variability is more appropriate than it would be to look for invariances, i.e., to examine the relationship between objects and the images they produce under changing illumination and find ways to compensate for its effect in an alignment style of approach.

The third source of variability has to do with the effect of varying context. A typical image often contains multiple objects that are next to each other, or partially occluding each other. If we attempt to compare the entire image (containing a familiar object) to the model representation of an object in question, then we are unlikely to have a match between the two. The problem of varying context is, therefore, a question of how the image representation of an object (say its contours) can be separated from the rest of the image before we have identified the object. The problem is difficult and is often referred to as the problem of "segmentation", "grouping" or "selection". In the context of achieving recognition the crucial question is whether the problem of context can be approached in a bottom-up manner, i.e., irrespective of the object to be recognized, or whether it requires top-down processes as well. We discuss this further in Section 1.4, but we can mention here that there is considerable empirical evidence, drawn from physiology and psychology, that the human visual system contains elaborate processes that perform segmentation prior to the subsequent recognition process.

The fourth source of variability has to do with objects changing their shape. These include objects with movable parts (such as the human body) and flexible objects (for example, a face where the changes in shape are induced by face expressions). This source of variability is geometrical, but unlike changing viewing positions, the geometric relation between objects and their images has less to do with issues of projective geometry and more to do with defining the space of admissible transformations in object space.

In this thesis we focus on the first two sources of variability, i.e., on geometric and photometric effects. The scope of the problem and its definition are discussed in the next section.

1.2 Scope of Recognition in this Work

The recognition problem we consider is that of identifying an image of an arbitrary individual 3D object. We allow the object to be viewed from arbitrary viewing positions, using the model of central projection, and to be illuminated by an arbitrary setting of light sources. We assume that the image of the object is already separated from the rest of the image, but may have missing parts (for example, as caused by occlusion).

We adopt the alignment methodology, which defines "success" as the ability to exactly re-construct the input image representation of the object (possibly viewed under novel viewing and illumination conditions) from the model representation of the object stored

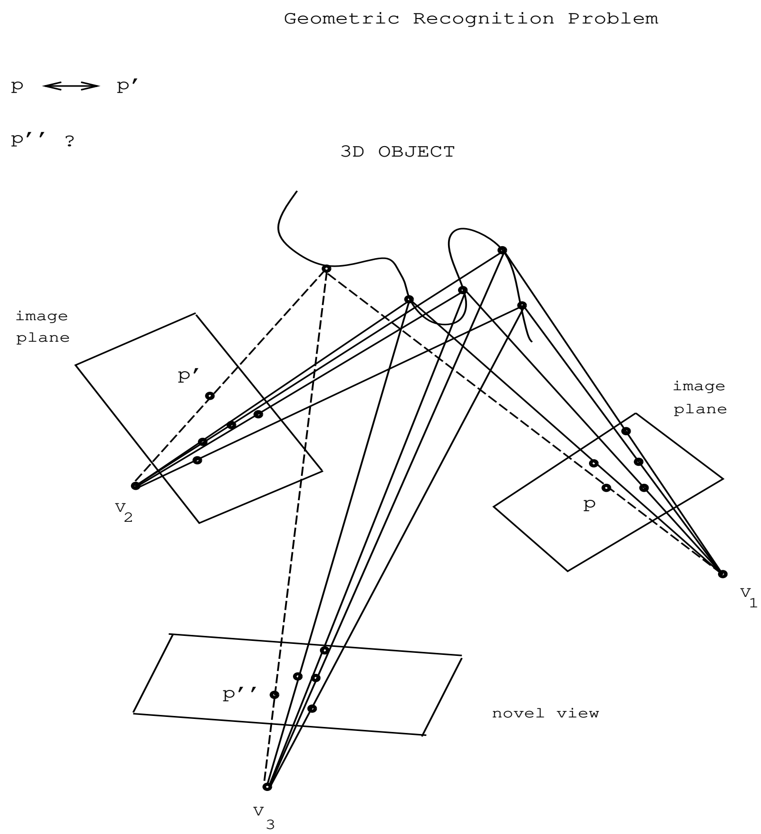


Figure 1.1: Illustrating the geometric alignment problem of recognition. Given a small number of corresponding points between the novel input view and the two model views, determine for any fifth point P, projecting onto p and p' in the two model views, the location p'' of its projection onto the novel image.

in memory. We assume low-level representations of both the object model and the input image. An object is represented by a small number of grey-level images, and the input image is represented by grey-levels, or points (edges, contours), or what we call "reduced" representations that are binary images made out of contours and sign-bits (such as those produced by thresholding the image, or by edge detection using a Laplacian of Gaussian operator).

The geometric and photometric components of the recognition problem can be treated independently of each other and then combined together into one recognition scheme. We therefore define the geometric and photometric problems as follows.

Definition 1 (Geometric Problem) *Given two projections (central, perspective, or parallel) of an arbitrary collection of points in 3D space (the object), then for any arbitrary*

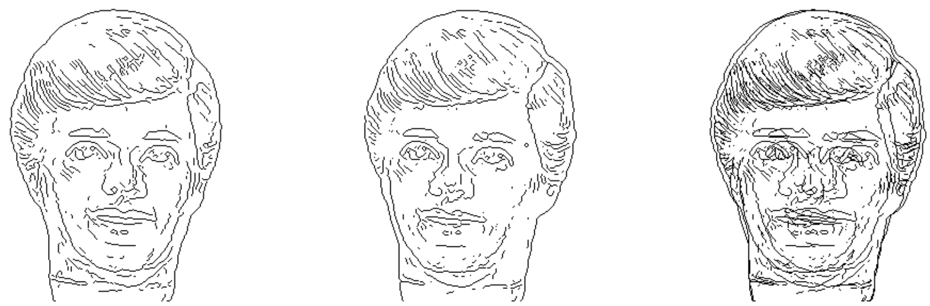


Figure 1.2: Demonstrating the effects of changing viewing position on the matching process. The difficulty of matching two different views can be illustrated by superimposing the two. One can see that, even for relatively small changes in viewing position, it could be very difficult to determine whether the two views come from the same face without first compensating for the effects of viewing transformation.

planar figure (novel image), determine whether it can be produced by a projection of the object.

The geometric problem (illustrated in Figure 1.1) assumes we can identify a small number of corresponding points across the different views of the same object, which we assume can be established by means of correlation (this is discussed in detail in Chapter 7). We note that even relatively small changes in viewing position between two images of the same object often create a real problem in matching the two against each other. Figure 1.2 illustrates this point by superimposing two edge images of a face separated by a relatively small rotation around the vertical axis. We see that it could be very difficult to determine whether they come from the same face without first compensating for the effects of viewing transformation.

Definition 2 (Photometric Problem) *We are given three images of an arbitrary convex matte surface. The images are taken under three different arbitrary settings of point light sources. For any arbitrary image determine whether the image can be produced by the surface under some illumination condition.*

The photometric problem is a question of how one can compensate for the effect of changing illumination by directly predicting the input signal, assuming that it came from the same surface that produced the model images. As with the geometric problem, this approach follows the alignment approach for recognition. The photometric problem also

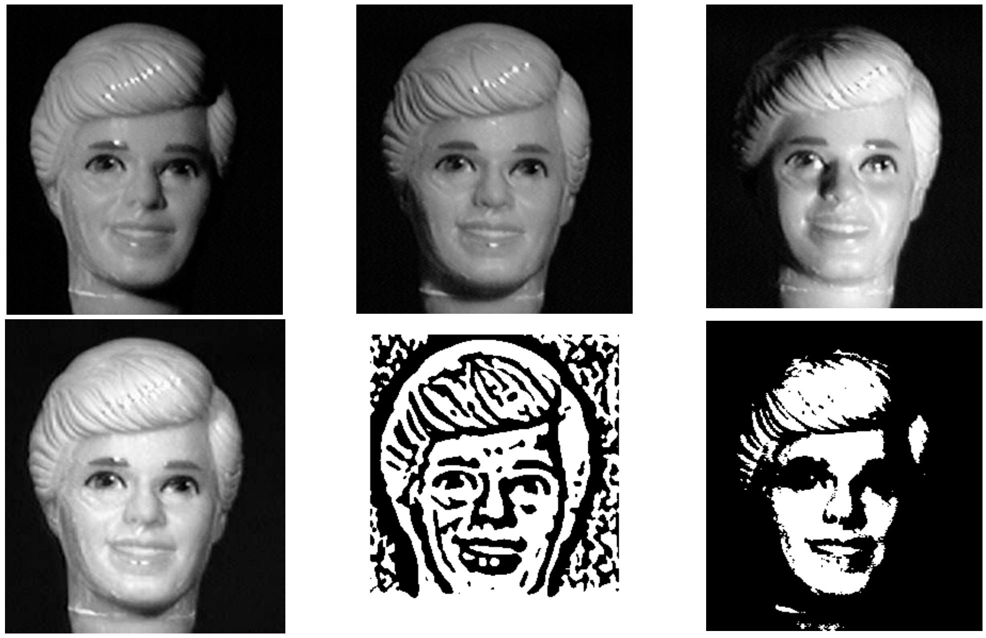


Figure 1.3: The photometric problem. The images of 'Ken' in the top row are taken from different illumination conditions (same viewing position). The images in the bottom row are various image representations of a novel image (novel illumination condition). The image on the left in the bottom row is the original novel image, the center image is the sign-bits of the Laplacian of Gaussian operator applied to the original image, and the image on the right is produced by thresholding the original image by some unspecified value.

raises the question of image representation. In other words, what is the minimal necessary information, extracted from the image, which will cancel the effects of changing illumination? The issue of representation is discussed in more detail in Chapter 4.

In practice, we work with an approximate version of the photometric problem by admitting non-convex, approximately matte surfaces illuminated under situations that produce cast-shadows and highlights. Figure 1.3 illustrates the photometric problem on the kind of objects and the input image representations, we work with in this thesis.

Note that the photometric problem assumes the surface is viewed from a fixed viewing position, and that the geometric problem assumes the views are taken under a fixed illumination condition (for a matte surface this means that the angle between the local surface orientation and the light sources remains fixed). The overall recognition problem that is addressed in this thesis is a combination of both problems.

Definition 3 (Combined Problem) *We assume we are given three model images of a*

3D matte object taken under different viewing positions and illumination conditions. For any input image, determine whether the image can be produced by the object from some viewing position and by some illumination condition.

1.3 Existing Approaches

Modern approaches for performing recognition fall into two classes: one is symbolic, in which the image undergoes a relatively elaborated data-driven process of extracting geometric parts and their spatial inter-relations, and which is then compared to a symbolic model representation of the object. The other approach is more pictorial and low-level, in which the data-driven component is relatively minor (to a degree of extracting contours, line approximation, corners and simple grouping criteria), and most of the efforts are placed at the level of recovering the model-to-image transformation and the model-to-image matching. Ullman (1986) refers to the former as "recognition by structural decomposition methods" and to the latter as "recognition by alignment methods" (for reviews see also Pinker 1984, Binford 1982).

The general idea behind structural decomposition methods is that the geometric source of variability, i.e. the effects of changing viewing positions, would be canceled over a wide range of viewing positions when the object is described in terms of a relatively small number of parts that are composed out of a library of shape primitives and that are also relatively simple and easy to compute from the image (Binford 1971, Marr & Nishihara 1978, Brooks 1981, Biederman 1985, Connell 1985, Hoffman & Richards 1986).

The main problem with the structural decomposition approach is that it mostly applies to simple objects with clearly identifiable parts. In the general case of complex objects (like a shoe or a face) it may be difficult to describe the object in terms of a relatively small set of geometric primitives that are at the same time common to many other objects as well (Ullman, 1986). The alternative of simplifying the part description to include edges and line segments may be unrewarding, because the resulting object description will be highly complex, which, in turn, may increase the susceptibility of the system to noise.

In the alignment approach the emphasis is placed not on the data-driven image analysis component but directly on the geometric relation between objects and their images. Object representations vary across alignment methods, but they all share the property that the representation is relatively low-level and does not require an elaborate data-driven component. The general idea behind the alignment approach involves a hypothesis-verification process. First a model-to-image transformation, called the alignment transformation, is recovered. The alignment transformation is then applied to the model in order to produce a synthesized image. The synthesized image is then compared to the actual input image for verification. The alignment transformation is the key component of this process and is responsible for compensating for the change in viewing position between the model and the input image. Such an approach was defined by Ullman (1986) and used also in Fischler & Bolles (1981), Lowe (1985), Faugeras & Hebert (1986), Huttenlocher & Ullman (1987), Thompson & Mundy (1987). Alignment methods differ in the following ways:

1. Object representation.
2. Recovery of alignment transformations. Types of recovery include:
   (a) A search over model-to-image correspondence space.
      i. Minimal alignment.
      ii. Constrained search over all possible correspondences.
      iii. Model pre-processing.
   (b) A search over transformation space.

Object representation is often based on a geometric structure that varies according to the information used to identify the object, the geometry used (metric versus non-metric), and the representation, i.e. whether it is explicit or embedded in the process for recovering the alignment transformation.

Some alignment methods identify the image of an object by reconstructing the 3D shape and comparing to it to the model (Douglass, 1981). Other alignment methods identify the image by predicting the appearance of the object and comparing it to the image (Huttenlocher & Ullman 1987, Lowe 1985). A 3D metric representation (i.e., one of relative depth) was used in Huttenlocher & Ullman (1987), and a 3D affine representation was implied in the work of Koenderink and Van Doorn (1991) on affine structure from two orthographic views. An implicit representation of affine structure was used by Ullman and Basri (1989) by modeling the object by two orthographic views in full correspondence. In other cases higher level representations are used by modeling the object by sets of identifiable features. An image is recognized if it contains a corresponding set of features (e.g., Fischler & Bolles 1981).

The methods for recovering the alignment transformation vary according to which space is searched over — model-to-image correspondence space, or transformation space. Some alignment methods determine the transformation by first identifying a small number of corresponding points between the image and the model (Fischler & Bolles 1981, Lowe 1987 for the perspective case, and Huttenlocher & Ullman 1987, Shoham & Ullman 1988, Ullman & Basri 1989 for the orthographic case). Consequently, in the case in which the alignment points are indistinguishable, the search space is over all possible tuples of points containing the minimal number of corresponding points required for recovering the alignment transformation. The correspondence problem can be constrained and the search space reduced if alignment points are not all indistinguishable, i.e., if they carry labels. For example, Huttenlocher and Ullman (1987) classify feature points into different types, such as corners and inflection points. Only points that carry the same label can match together, therefore reducing the search space.

Other methods that search over the image-to-model correspondence space search over the space of all possible correspondences between the set of image features and the set of model features. The search space is reduced by defining constraints, often pairwise constraints, that follow certain "perceptual organization" or "grouping" rules, such as proximity of features, connectivity, collinearity and parallelism. Search over all possible correspondences has the advantage of not assuming that the image of the object is isolated from the rest of the image. Therefore, these methods attempt to deal with the varying context source of variability in addition to the geometric source. Because of the grouping rules that are used for managing the search, these methods are often limited to recognizing images of relatively simple objects such as polyhedrons and simple machine parts (Roberts 1965, Davis 1979, Bolles & Cain 1982, Grimson & Lozano-Pérez 1984, Faugeras & Hebert 1986, Van Hove 1987, Lowe 1985,1987).

Another method for reducing the search over the model-to-image correspondence space is the "geometric hashing" method introduced by Lamdan, Schwartz & Wolfson (1988). The idea is similar to minimal alignment, but with a pre-processing stage in which multiple copies of the object model, one for each tuple of alignment points, is stored in a table. This method has the advantage of there being no need to establish a model-to-image correspondence. Furthermore, more than one object model can be stored in the table. The latter property implies that the search over different objects can be done in parallel rather than in serial, as with the other alignment methods mentioned above. The main problem with the geometric hashing method, however, is that it is most suitable to planar objects or to 3D polyhedrons and is more sensitive to noise than minimal alignment without the pre-processing stage (Grimson, Huttenlocher & Jacobs, 1991). Jacobs (1992) proposed another model pre-processing method that can deal with 3D objects, but because of grouping rules employed during model acquisition, it is limited to relatively simple objects.

Another set of alignment methods search over the transformation space rather than over correspondence space. The best known example of this approach is the generalized Hough transform introduced by Ballard (1981). In the Hough transformation method every tuple of alignment points votes for the transformation it specifies (the parameters of the viewing transformation). The transformation that is supported by the largest number of tuples is selected and then verified by matching the transformed model with the actual image. The method is sensitive to noise because the voting table must sample a six dimensional space (six degrees of freedom for defining a rigid viewing transformation) and because of quantization problems (Grimson & Huttenlocher 1988). Other examples of the transformation space search approach include the deformable templates method (Yuille, Cohen & Hallinan 1989), the local search in transformation space by maximizing a probability density function (Wells 1992), search by transformation clustering (Thompson & Mundy, 1987), search by transformation sampling (Cass 1988), and combined correspondence and transformation space search (Cass 1992, Breuel 1992).

1.4 Relationship to Human Vision

Of particular interest to any theory in machine vision is to find some biological evidence at the physiological or psychophysical level for the analytic problems identified by the theory and for the approach by which those problems should be solved. The best known examples of a successful match between an analytic model and biological data are in the field of edge detection (Hubel & Wiesel 1962, Marr & Hildreth 1980) and the measurement of retinal motion (Hildreth, 1984). Visual recognition, on the other hand, involves a considerable degree of abstraction which precludes direct inspection of the lower substrate of processing levels.

In the sections below we explore the analytic aspects of the recognition problem (which were described in Section 1.1 in terms of sources of variabilities) from the standpoint of available biological data. We will focus on data concerning the extent of low-level processing that is done prior to recognition, and the role of geometric and photometric cues in human visual processing.

1.4.1 Recognition and the Problem of Varying Context

The problem of varying context raises the question of whether the problem of recognition can be isolated and treated independently of the data-driven segmentation process, or whether the two are strongly coupled. It appears that in some cases in human vision the processes for performing grouping and segmentation cannot be isolated from the recognition process. In some well known examples, such as R.C. James' image of a Dalmation dog (see, Marr 1982), it appears unlikely that the image of the object can be separated from the rest of the image based on image properties alone and, therefore, some knowledge about the specific class of objects is required to interpret the image.

Human vision, however, appears also to contain relatively elaborate processes that perform grouping and segmentation solely on a data-driven basis independent of subsequent recognition processes. For example, Kinsbourne and Warrington (1962, cited in Farah 1990) report that patients with lessions in the left inferior temporo-occipital region are generally able to recognize single objects, but do poorly when more than one object is present in the scene. Another line of evidence comes from displays containing occlusions. The occluding stimuli, when made explicit, seem to stimulate an automatic 'grouping' process that groups together different parts of the same object (Nakayama, Shimojo & Silverman, 1989). The third line of evidence comes from 'saliency' displays in which structures, not necessarily recognizable ones, are shown against a complex background. Some examples are shown in Figure 1.4. In these displays, the figure-like structures seem to be detected immediately despite the lack of any apparent local distinguishing cues, such as local orientation, contrast and curvature (Shashua & Ullman, 1988).

1.4.2 Geometry Related Issues in Human Vision

The use of geometric information for visual analysis of shape is not limited to the task of object recognition. We use shape and spatial information for manipulating objects, for planning and following a path, and for performing movements in the environments in general. The use of geometric information subserving recognition and motor interactions are therefore two related but separate issues. Furthermore, even within the context of performing recognition, geometric information may be used differently for the task of identifying individual objects and for classifying an object as belonging to a particular class of objects. Because of the diverse application of geometric information in visual analysis, and the apparent difficulty in decoupling these issues at the experimental level, most of

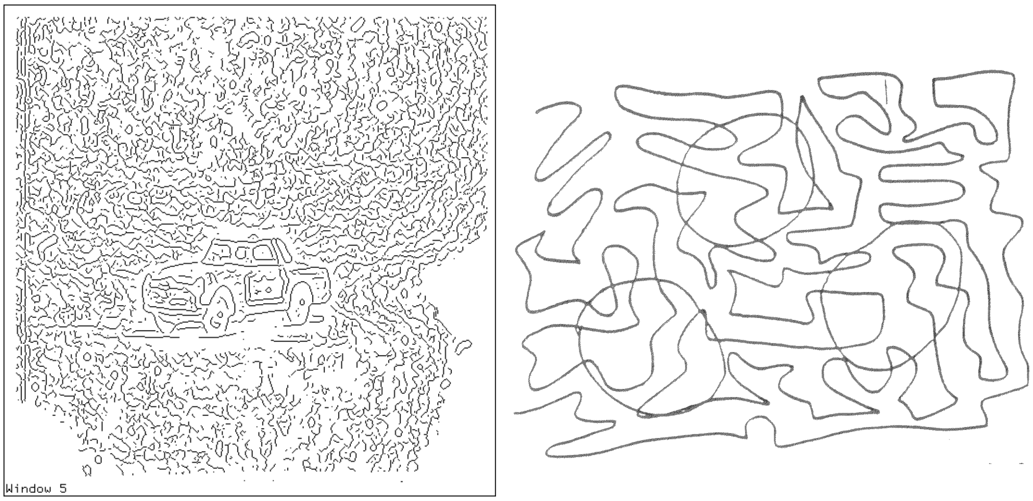


Figure 1.4: Structural-saliency displays. The figure like structures seem to 'pop-up' from the display, despite the lack of any apparent local distinguishing cues, such as local orientation, contrast and curvature (Shashua and Ullman, 1988). The figure on the right originally appeared in (Mahoney, 1986)

the empirical data available on human visual recognition are not conclusive in ruling out competing theories, but rather serve to support some of the existing ones. The general outline of using an alignment transformation prior to matching an image to a model, the use of viewer-centered model representations of shape, the use of pictorial information at the level of model-to-image matching, and the use of non-metric structure representation, appear to agree with several observations on human vision. We discuss below some of the empirical data that tend to support these aspects of the alignment approach and the use of non-metric representations.

Orientation Alignment in Human Vision

The empirical evidence related to the possible role of an alignment transformation occurring during the recognition process comes from studies on "mental rotations". These studies establish the existence of recognition latencies for matching shapes that differ in their depth orientation, with latencies increasing directly with degree of disparity in orientation (Jolicœur 1985, Shepard & Metzler 1988, Tarr & Pinker 1989, Edelman & Bülthoff 1990). These findings suggest that during recognition the orientation of the viewed object is brought into alignment with its corresponding stored model.

Viewer-centered Model Representations

The second aspect of the alignment approach is that models are viewer-centered, i.e., the model describes the object as seen from a particular viewing position, or from a particular restricted range of viewing positions (in case a number of 2D images are used for model representation, as opposed to a 3D representation). The possible use of a viewer-centered model representation is supported by studies showing that recognition is severely impaired when the disparity in orientation between the learned object and the viewed object becomes too large (Rock, DiVita & Barbeito 1981, Rock & DiVita 1987). Edelman & Bülthoff (1990) also show that performance in recognition is best for novel views that are in between the learned views, and that performance degrades with increasing angular separation between the novel and learned views.

There is also some physiological evidence supporting the notion of viewer-centered representations. Recordings from face-sensitive neurons in the macaque's STS suggest that memory representations for faces are viewer-centered , and that each representation is usually view-insensitive, covering a rather wide range of orientations in space (Perret, Smith, Potter, Mistlin, Head, Milner and Jeeves, 1985).

Pictorial Matching

Another aspect of the alignment approach is that the match between the image and the model representation stored in memory is performed at a low level of matching pictorial descriptions, or template matching, rather than employing symbolic descriptions. Empirical evidence suggests that a pictorial comparison between an image and a model is a possible in some cases. Palmer (1978) conducted experiments that show that in tasks of simultaneous comparison (two figures presented simultaneously) subjects tend to use structural and abstract features such as closure and connectivity. In contrast, in sequential comparison tests (essentially a recognition test) the main determinant is the degree of pictorial overlap.

Non-metric Structure

Another aspect of our recognition approach is that the structure representation of objects is not necessarily metric. Non-metric representations imply either a more flexible camera model (central projection instead of perspective projection), or equivalently, that objects are allowed to undergo non-metric transformations, such as stretch and shear. There is only limited empirical evidence regarding what form of structure information is used for representing models. The existing empirical data, though not specific to recognition, suggest that the kind of geometric information employed by human vision is not necessarily metric.

Luneburg (1947) and Schelling (1956), (see also Julesz 1971) have argued that for a given fixation point the measurements of binocular vision are non-metric, and this is because as we change eye convergence from one target to the next, the perceived distance of the previous target does not seem to change. Cutting (1986) offers another reason why non-metric representations may be preferred over metric ones by referring to *La Gournerie's paradox*: Visual interpretation of 3D objects from pictures appears to be robust even in situations in which the pictures are viewed from the side (see also Kubovy 1986, Jacobs 1992). This observation implies that central projection may be more appropriate than perspective projection when modeling the geometric relation between objects and their images (see Section 1.6). Recently, Todd and Bressan (1990) have suggested using affine representations of structure based on psychophysical experiments on human subjects. Their experiments suggest that affine properties play an important role in the perception of kinetic depth displays, even in cases where the number of views presented to the subjects were more than sufficient to recover metric properties.

1.4.3 Issues of Photometry in Human Vision

The problem of varying illumination conditions, or the photometric problem as we refer to it here, raises the question of whether the problem can be isolated and dealt with independently of subsequent recognition processes, or whether it is coupled with the recognition process.

It appears that in some cases in human vision the effects of illumination are factored out at a relatively early stage of visual processing and independently of subsequent recognition processes. A well known example is the phenomenon of lightness and color constancy. In human vision the color of an object, or its greyness, is determined primarily by it's reflectance curve, not by the actual wavelengths that reach the observer's eye. This property of the visual system is not completely robust as it is known, for example, that fluorescent lighting alters our perception of colors (Helson, Judd & Wilson, 1956). Nevertheless, this property appears to suggest that illumination is being factored out at an early stage prior to recognition. Early experiments that were used to demonstrate this used simple displays such as a planar ensemble of rectangular color patches, named after Mondrians'

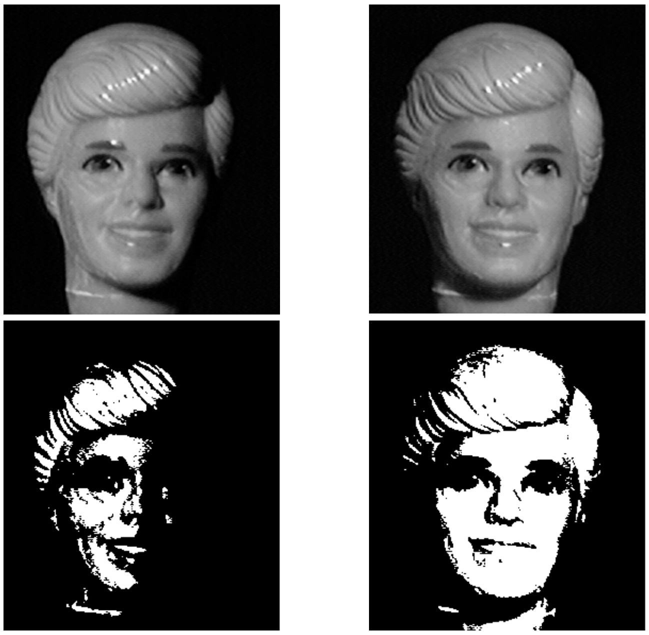


Figure 1.5: Images of 'Ken' taken from different illumination conditions followed by a thresholding operation. The recognizability of the thresholded images suggests that some knowledge about objects is required in order to factor out the illumination, and specifically that the image we are looking at is an image of a face.

paintings, or comparisons between Munsel chips (Land & McCann, 1971). More recent psychophysical experiments demonstrated the effect of 3D structure on the perception of color and lightness (Gilchrist 1979, Knill & Kersten 1991). These experiments show that the perception of lightness changes with the perceived shape of the object. The objects that were used for these experiments are relatively simple, such as cylinders, polyhedrons and so forth. It is therefore conceivable that the 3D structure of the object displayed in these kinds of experiments can be re-constructed on the basis of image properties alone after which illumination effects can be factored out.

Human vision, however, appears also to contain processes that factor out the effect of illumination during the recognition process. In other words, the image and the model are coupled together early on in the stages of visual processing. Consider, for example, the images displayed in Figure 1.5. The images are of a 'Ken' doll lit by two different illumination conditions, and thresholded by an arbitrary value. The thresholded images appear to be recognizable, at least in the sense that one can clearly identify the image as containing a face. Because the appearance of the thresholded images critically rely on the illumination conditions, it appears unlikely that recognition in this case is based on the input properties alone. Some knowledge about objects (specifically that we are looking at the image of a face) may be required in order to factor out the illumination. The issue of the recognizibilty of reduced image representations, and the issue of image representation in general, is discussed in more detail in Chapter 4.

1.5 Overview of Thesis Content and Technical Contributions

This thesis is organized in three parts (see Figure 1.6). The first part of the thesis (Chapter 2 and 3) considers the geometric relationship between 3D objects and the images they produce under central projection. The results established in this part have direct contributions to the geometric problem of recognition and to the representation and recovery of relative structure under the most general conditions (all projection models are treated alike, internal camera calibration is not necessary).

The second part of this study (Chapters 4,5 and 6) considers the photometric problem of recognition. We consider the problem of factoring out the illumination during the recognition process in a model-based approach. The model-based approach proceeds by establishing a connection between different images of the same object under changing illumination. This connection provides an algorithm by which a novel image can be reproduced by three model images of the object.

The third part of the thesis (Chapters 7 and 8) considers the correspondence problem and the combined recognition problem. This part is distinct from the other two parts of the thesis because here we begin to consider both sources of information together. We show that the correspondence problem, which is a necessary component in building models of objects, can be approached by combining affine or projective geometry and optical flow into a single framework. We then consider the problem of achieving recognition under both sources of variability — changing viewing positions and illumination conditions — occurring simultaneously. The issues and contributions made in each part of the thesis is described in more detail in the following sections.

1.5.1 Part I: Geometry, Recognition and SFM

The geometric relationship between 3D space and image space is a topic of interest in recognition and in structure from motion. Most of the mathematical problems in this topic are well understood when it comes to parallel and orthographic projections. However,

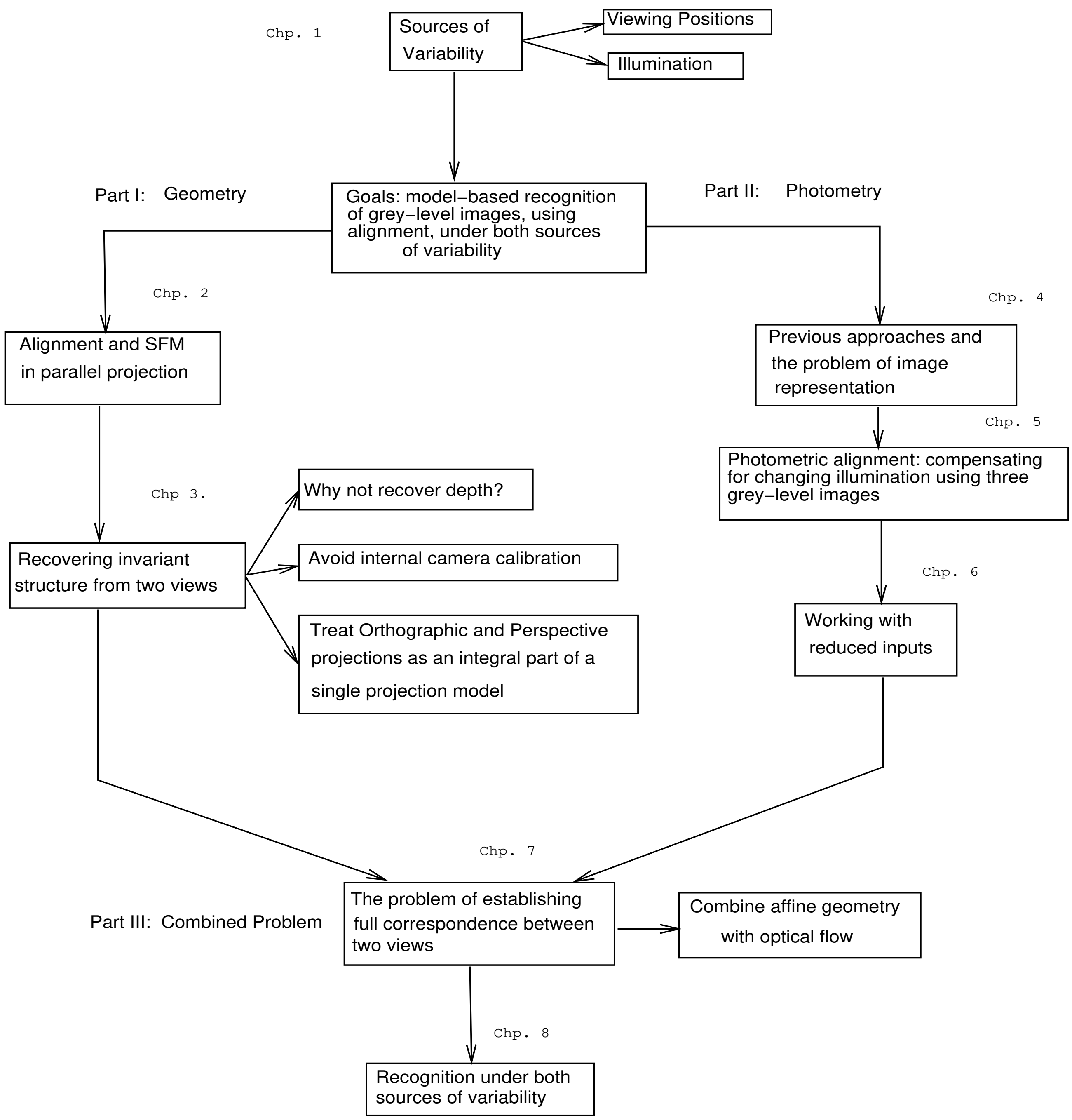


Figure 1.6: Graphic Roadmap of the thesis. The thesis is organized in three parts: the geometric part deals with the geometric relation between 3D space and image space under central projection. The photometric part deals with the problem of compensating for the effects of changing illumination in matte surfaces using a model-based approach. The third part combines geometric an photometric sources of information to solve the correspondence problem, and the combined recognition problem.

many hard mathematical problems remain when we consider the more general models of projection — central and perspective projection.

We argue that at the level of problem definition, three major issues must be addressed before one attempts to propose solutions. The first issue is why there should be a distinction made between orthographic and perspective projections. Virtually all previous methods assume either one or the other, and often perform best when the field of view is either very wide (strong perspective distortions) or very narrow. The second question is whether internal camera calibration is really necessary. Thirdly, we analyze whether structure needs to be metric, and if not, what kind of non-metric structure should be sought.

Our major contribution is to propose a framework that provides an adequate solution to these problems requiring only simple linear computations that is useful also for purposes of recognition under the alignment approach. There are two components to this framework: (i) the model of central projection is used instead of perspective projection, (ii) the non-metric structure of the object is defined in a way that does not implicate the center of projection, thereby allowing the center of projection to be any point in projective space, including the case of an ideal point (parallel projection).

1.5.2 Part II: The Photometric Problem

Chapter 4 addresses the photometric problem both from a practical point of view and from empirical observations of human vision. Two central questions are addressed in this chapter: first, is there a need for a model-based approach for dealing with the effects of illumination in recognition? Second, what are the limits on image information in order to make that process work? The evidence we look at suggest, first and foremost, that image properties alone do not appear to be sufficient for a complete solution, and secondly, that the process responsible for factoring out the illumination during the recognition process appears to require more than just contour information, but just slightly more.

In Chapter 5 we introduce the basic method, we call photometric alignment, for compensating for the effects of illumination during recognition. The method in based on a result that that three images of the surface provide a basis that spans all other images of the surface (same viewing position, but changing illumination conditions). The photometric problem of recognition is, therefore, reduced to the problem of determining the linear coefficients. Chapter 6 extends the basic method to deal with situations of recognition from reduced image representations. The computational results introduced in this chapter appear to agree with the empirical observation made in Chapter 4 that sign-bits appear to be sufficient for visual interpretation, whereas edges alone do not.

The conclusion is therefore, that with regard to the computational aspect of the photometric problem, three model images of the object can be used to reproduce any novel image of the object, even in the case where only a "reduced" input is provided to the recognition system. The minimal reduction that is still sufficient for purposes of recognition is not to the level of edges, but to the level of edges and their sign-bits (includes the case of thresholded images).

1.5.3 Part III: Combining Geometric and Photometric Sources of Information

This part of the thesis deals with the problem of achieving full correspondence between two images taken from different viewing positions and the combined recognition problem, i.e., recognition under changing illumination and viewing positions.

The correspondence problem is a critical component of the alignment-based approach. Both the geometric and photometric components assume that the model of the object is represented by a small number of images for which all correspondences are known. Chapter 7 deals with the correspondence problem. The approach to this problem is to combine both the geometric and photometric sources of information in order to fully determine all correspondences between two views of a rigid object. The main result is that a small number of known correspondences, together with the observed temporal and spatial derivatives of image intensities everywhere else, are sufficient to uniquely determine the correspondences between all image points in both views.

The combined recognition problem is addressed in Chapter 8. The first two parts of this thesis dealt with each source of variability separately and independently because the effects of changing geometry and changing illumination are decoupled when dealing with matte surfaces. We can, therefore, combine the results that were derived in Part I, and part II of the thesis in order to deal with the combined problem, i.e., recognize novel images of the object taken from novel viewing positions and novel illumination conditions.

1.6 Projection Models, Camera Models and General Notations

The geometric relation between 3D space and image space depends, first and foremost, on what assumptions are being made on the way the world is projected onto the image plane.

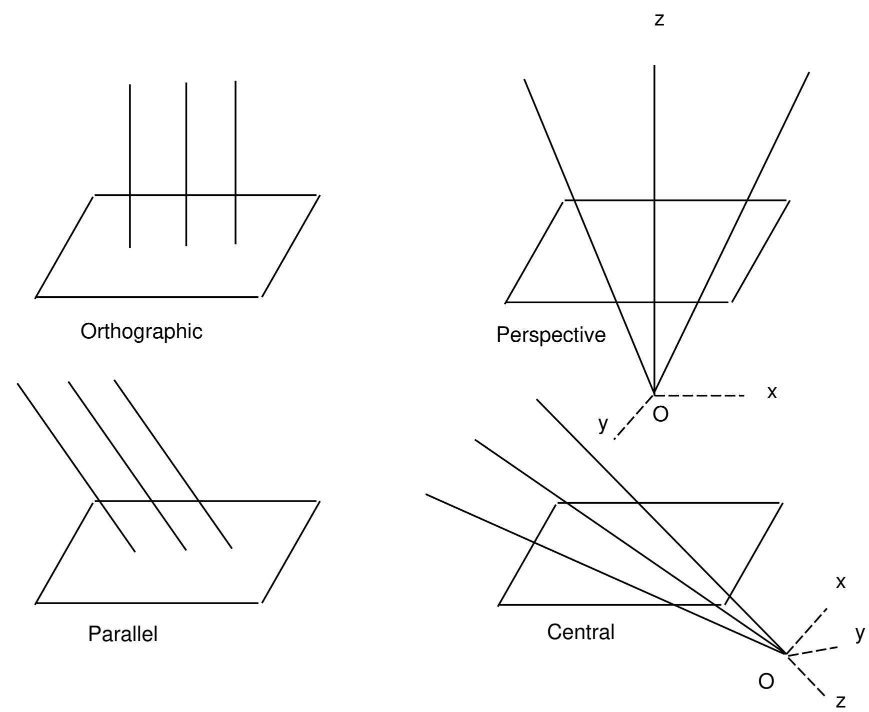


Figure 1.7: Projection models: orthographic, parallel, perspective and central.

We distinguish between four models of projection: orthographic, perspective, parallel and central. The first two models are mostly used in computer vision, while the latter two are used in models of higher geometries of the plane.

Figure 1.7 illustrates these four projection models. To project a collection of points in 3D space onto a plane via a point O, we draw lines from O to the different 3D points. The image is produced by the intersection with a plane (the image plane) which is not coplanar with O. This is known as central projection, and O is known as the center of projection (COP). Parallel projection is the same thing but with the condition that the center of projection O is at infinity ("ideal" point in 3D projective space). In parallel projection, therefore, the rays are all parallel to each other.

The perspective projection model is similar to central projection but with several additional conditions. First, perspective projection has a distinguishable ray known as the optical axis. Second, the optical axis not only intersects the image plane at a fixed point (known as the principal point), but is also perpendicular to the image plane. Third, the distance of the image plane from the center of projection along the optical axis is fixed and is known as the focal length. A more convenient way to describe perspective projection is to assume a coordinate frame with its origin at the center of projection, its z axis aligned with the optical axis, and its xy plane parallel to the image plane. In perspective projection, therefore, the location of the image plane with respect to this coordinate frame is assumed to be known — whereas in central projection the location of the image plane with respect to the coordinate frame is arbitrary and unknown. In other words, the change of coordinates between two perspective views is assumed to be rigid, i.e., translation of the COP followed by a rotation of the coordinate axes, whereas the change in coordinates between two central projections of the same object is composed of a rigid motion followed by an arbitrary projective transformation of the image plane. In particular, this means that we can have an affine change of coordinates (the xyz frame undergoes an arbitrary linear transformation in space, rather than only rotations), and in addition we can take a projection of a projection (taking a view of the image plane).

The orthographic projection model is similar to parallel projection, but with the condition that the image plane is perpendicular to the projecting rays (perspective projection in which the center of projection is an ideal point). In computer vision, uniform scale is often added to the orthographic model in order to model changes in size due to the distance between the image plane and the object. The scale extended orthographic model is then often referred to as "scaled orthographic projection" or "weak perspective". For reasons of simplicity, we will continue to refer to this as orthographic projection, with the understanding that uniform scale is included.

1.6.1 Camera Models

The perspective projection model describes an internally calibrated camera. The location of the principle point, the focal length and the true angle that the image plane makes with the optical axis of the particular camera in use, are often known as internal camera parameters. Perspective projection is, therefore, an accurate model of the way the world projects onto film — provided that we have full knowledge of internal camera parameters. We refer to this imaging model as a calibrated camera model, or perspective projection with a calibrated camera, or a rigid camera (rigidity comes from the observation that knowledge of internal parameters is inter-changeable with assuming that the world is rigid).

The viewing transformation with a rigid camera is, therefore, composed of camera translation and rotation of the camera coordinate frame around the new location of the center of projection. This is often referred to as the six parameter motion of the camera.

The six parameters include two for translation (magnitude of translation cannot be recovered from 2D observations alone), three for specifying the rotation axis in space, and one for the angle of rotation.

The model of central projection describes an uncalibrated camera. In this case, the camera coordinate frame can undergo any arbitrary linear transformation in space (around the new location of the COP), rather than only rotations. This means that the location of the principle point is not fixed as the camera changes its position. In addition, we allow the views to undergo arbitrary projective transformations of the plane, which is equivalent of taking arbitrary central projections of central projections of the object (as when looking at a picture of an object). We refer to this camera model as an uncalibrated camera or a non-rigid camera. A non-rigid camera has, therefore, the advantage of not requiring prior knowledge of internal parameters, allows us to take pictures of pictures of 3D objects, which taken together means that only non-metric world properties can be recovered from the 3D scene.

Orthographic and parallel projections are an approximation to the rigid and non-rigid camera models, respectively. The approximation holds under conditions of small field of view (objects are relatively far away from the camera) of objects which are only moderately extended in depth. Parallel projection is equivalent of taking orthographic views followed by an arbitrary affine transformation of the plane. Orthographic views followed by affine transformations of the plane are in turn equivalent to having orthographic views and allowing the object to undergo arbitrary affine transformations in space (David Jacobs, personal communication).

In this thesis we address the central, perspective and parallel models of projection. We therefore address the geometric problem of recognition and the problem of recovering relative structure from two views in situations where calibration is unknown as well as in situations where calibration is assumed.

1.6.2 General Notations

We denote object points in capital letters and image points in small letters. If P denotes an object point in 3D space, p, p', p'' denote its projections onto the first, second and novel projections, respectively. We treat image points as rays (homogeneous coordinates) in 3D space, and refer to the notation $p = (x, y, 1)$ as the standard representation of the image plane. We note that the true coordinates of the image plane are related to the standard representation by means of a projective transformation of the plane. In case we deal with central projection, all representations of image coordinates are allowed, and therefore, without loss of generality, we work with the standard representation (more on that in Appendix A).

Part I

Geometry: Recognition and Structure from Motion

Non-metric Structure and Alignment from two Parallel Projected Views

Chapter 2

The geometric problem we consider in this study is the recognition, via alignment, of a novel view given a small number of corresponding points with two model views of the same 3D object. In other words, given the image locations of a small number of object points in the two model views and the novel view, we would like to be able to determine the image locations of all other points that project onto the novel view. We refer to this problem as "re-projection" for we wish to re-project the model of the object (represented by two views) onto any arbitrary viewing position. Once we have re-projected the model, we can compare the re-projected view and the novel view and expect to have a match when the three views are projected from the same set of 3D points.

This chapter sets the stage by first focusing on a relatively simple domain resulting from assuming parallel projection. We start with the more simple problem first primarily because we argue that the general case is not much different and naturally extends from it — provided we look at the simple case from the proper angle. Most of the material covered in this chapter is related to the work of Koenderink and Van Doorn (1991) on affine structure from motion, and the work of Ullman and Basri (1989) on the linear combination of views. The main points covered here include the following:

- Non-metric structure is introduced in two forms: affine coordinates, and affine structure.

- Re-projection can be achieved by recovering the epipolar geometry and the non-metric structure of the object. Epipolar geometry alone is also sufficient in most cases, but generally provides only a weak solution to the problem of re-projection.

- The linear combination of views can be derived from the affine structure result.

2.1 Overview

In the context of structure from motion (SFM) and recognition under the assumption of orthographic projection, it is known that three views are necessary for recovering metric structure (Ullman 1979, Huang & Lee 1989, Aloimonos & Brown 1989, Tomasi & Kanade 1991, Weinshall 1992) and, similarly, three model views are necessary for achieving recognition of rigid objects (Ullman & Basri, 1989). The conclusion, therefore, is that structure and/or recognition from two model views is governed by affine geometry rather than Euclidean geometry. The fact that this is possible is due to Ullman and Basri (1989) and Koenderink and Van Doorn (1991).

One way to view non-metric (in this case affine) structure is by coordinates. Instead of recovering the coordinates with respect to the camera coordinate frame (the image plane coincides with the xy plane, and the optical axis coincides with the z axis) we may consider recovering the coordinates of object points with respect to a coordinate frame defined by four non-coplanar object points whose positions in space are unknown. We therefore cannot measure distances in space (the angles between the affine axes are not preserved under affine transformations), and the coordinates are not absolute because they depend on the choice of the four points. Nevertheless, with affine coordinates we get some feeling of structure (we can tell the shape of the object up to an unknown stretch and shear) and can predict the appearance of the object as seen from any other novel position under parallel projection.

Another way to represent non-metric structure is to define a measurement of the point of interest P with respect to the four points, such that the measurement remains fixed under parallel projection. For instance, we may consider three of the four points defining a plane and the invariant measurement may be some form of relative deviation of P from the plane. Koenderink and Van Doorn (1991) refer to this as "affine structure" and show that it is equivalent to the third affine coordinate of P. In a way Similar to the case of the affine coordinate representation, we can predict novel views of the object by carrying only one number (the affine structure invariant) instead of three numbers (affine coordinates) for each point.

The representation of structure by affine coordinates or by affine structure are equivalent as we can convert one to the other. The difference is in concept. We shall see later that the difference is significant when it comes to the general case of central projection.

Another strategy for achieving re-projection, described by Ullman and Basri (1989), is by a result which appears to be orthogonal to the affine structure result in the sense that no structure is explicitly involved in the process of re-projection. Ullman and Basri show that image coordinates of corresponding points in three views (two model views and the novel view) are linearly related: the image coordinates of the novel view can be obtained by linearly combining the image coordinates of the two model views. We will establish a connection between the two results by showing that the linear combination result can be derived from the affine structure result.

2.2 Affine Coordinates from two Views

Let O, P_1, P_2, P_3 be four non-coplanar object points, referred to as reference points, and let O', P'_1, P'_2, P'_3 be the coordinates of the reference points from the second camera position. Let b_1, b_2, b_3 be the affine coordinates of an object point of interest P with respect to the basis OP_1, OP_2, OP_3, i.e.,

$$OP = \sum_{j=1}^{3} b_j(OP_j),$$

where the OP denotes the vector from O to P. Under parallel projection the viewing transformation between the two cameras can be represented by an arbitrary affine transformation, i.e., $O'P' = T(OP)$ for some linear transformation T. Therefore, the coordinates b_1, b_2, b_3 of P remain fixed under the viewing transformation, i.e.,

$$O'P' = \sum_{j=1}^{3} b_j(O'P'_j).$$

Since depth is lost under parallel projection, we have a similar relation in image coordinates:

$$op = \sum_{j=1}^{3} b_j(op_j) \tag{2.1}$$

$$o'p' = \sum_{j=1}^{3} b_j(o'p'_j). \tag{2.2}$$

Given the corresponding points p, p' (in image coordinates), the two formulas 2.1,2.2 provide four equations for solving for the three affine coordinates associated with the object point P that projects to the points p, p'. Furthermore, since the affine coordinates are fixed for all viewing transformations, we can predict the location p'' on a novel view by first recovering the affine coordinates from the two model views and then substituting them in the following formula:

$$o''p'' = \sum_{j=1}^{3} b_j(o''p''_j).$$

We have, therefore, a method for recovering affine coordinates from two views and a method for achieving re-projection given two model views and four corresponding points across the three views.

A more concise representation of non-metric structure and a more direct re-projection method can be derived if, instead of recovering affine coordinates, we define and recover an affine invariant. The affine invariant discussed next is what Koenderink and Van Doorn (1991) call "affine structure", which turns out to be simply b_3. More importantly, however, is the concept of describing structure as an invariant measurement with respect to a geometric construction defined by the reference points. We will use this concept later on when we deal with central projection.

2.3 Affine Structure: Koenderink and Van Doorn's Version

We can view the four reference points O, P_1, P_2, P_3 in space as composed of a reference plane, defined by O, P_1, P_2, and a reference point P_3 not coplanar with the reference plane. The fundamental theorem in affine geometry of the plane states that correspondences of three points uniquely determine all other correspondences of the plane (see Appendix A.1 for more details). In other words, from the observed correspondences in both views $o \longleftrightarrow o', p_1 \longleftrightarrow p'_1, p_2 \longleftrightarrow p'_2$ we can recover a transformation $T[]$ (affine transformation of the plane in non-homogeneous coordinates) that accounts for all correspondences induced by the reference plane. For example, for any point $\tilde{P}$ coplanar with the reference plane projecting onto $p, \tilde{p}'$ in our two views, then $\tilde{p}' = T[p]$.

Let P be an arbitrary point in the scene projecting onto p, p' on the two image planes. Let $\tilde{P}$ be the projection of P onto the reference plane along the ray towards the first image plane, and let $\tilde{p}'$ be the projection of $\tilde{P}$ onto the second image plane (p' and $\tilde{p}'$ coincide if P is on the reference plane). Note that the location of $\tilde{p}'$ is is determined by $T[p]$. Using a simple geometric drawing, the affine structure invariant is derived as follows.

Consider Figure 2.1. The projections of the fourth reference point P_3 and the point P

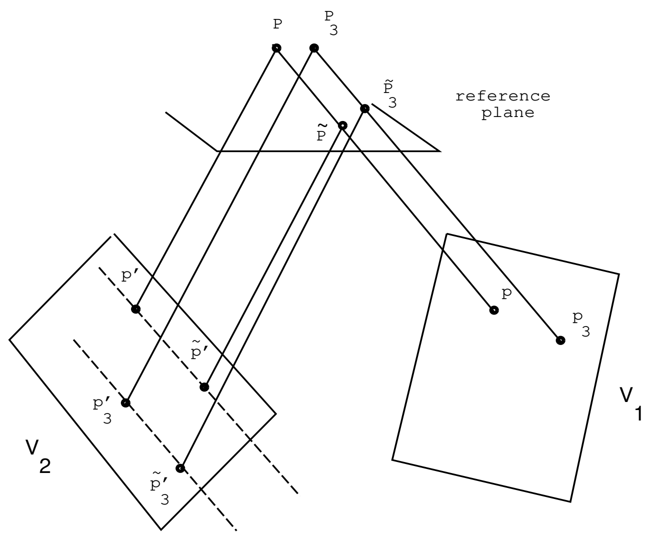


Figure 2.1: Koenderink and Van Doorn's Affine Structure.

form two similar trapezoids: $P\tilde{P}p'\tilde{p}'$ and $P_3\tilde{P}_3p'_3\tilde{p}'_3$. From similarity of trapezoids we have,

$$\gamma_p = \frac{|P - \tilde{P}|}{|P_3 - \tilde{P}_3|} = \frac{|p' - \tilde{p}'|}{|p'_3 - \tilde{p}'_3|}.$$

Since the motion of the camera consists of translation and rotation of the image plane in space and possible change of angle with repect to the projecting rays, γ_p is invariant to camera motion under parallel projection.

The affine structure γ_p is a measure of affine shape, just as affine coordinates were a representation of affine shape. With affine structure we can describe the location of a point in space relative to a reference plane whose orientation in space is unknown, therefore, we can tell what the object is up to an unknown shear and stretch.

We can also achieve re-projection onto a novel view by first recovering γ_p from the known correspondences between the two model views. The location of p'' is then determined by substitution of γ_p in the equation

$$\gamma_p = \frac{|p'' - \tilde{p}''|}{|p''_3 - \tilde{p}''_3|}.$$

Finally, the relation between affine coordinates and affine structure is simply $\gamma_p = b_3$, or in other words γ_p is a measure of "affine depth" (if we consider the analogy between the third affine coordinate and the z coordinate in metric space). This can be shown as follows:

Let A and $\boldsymbol{v}$ be the 2D affine transformation that aligns o, p_1, p_2 with o', p'_1, p'_2, i.e., $o' = Ao + \boldsymbol{v}$ and $p'_j = Ap_j + \boldsymbol{v},\ j = 1, 2$. By subtracting the first equation from the other two we get:

$$o'p'_1 = A(op_1)$$
$$o'p'_2 = A(op_2).$$

We can, therefore, see that A is the matrix $[o'p'_1, o'p'_2][op_1, op_2]^{-1}$. From the two formulas 2.1,2.2 we obtain the following result:

$$\begin{aligned} o'p' &= \sum_{j=1}^{3} b_j(o'p'_j) = b_1A(op_1) + b_2A(op_2) + b_3o'p'_3 + b_3A(op_3) - b_3A(op_3) \\ &= A(op) + b_3(o'p'_3 - A(op_3)). \end{aligned}$$

By substituting $\boldsymbol{v} = o' - Bo$ we get:

$$p' = Ap + \boldsymbol{v} + b_3(p'_3 - Ap_3 - \boldsymbol{v}).$$

Finally, by noting that $\tilde{p}' = Ap + \boldsymbol{v}$ we get:

$$b_3 = \frac{|p' - \tilde{p}'|}{|p'_3 - \tilde{p}'_3|}.$$

2.4 Epipolar Geometry and Recognition

In the course of deriving affine structure, we have also obtained the epipolar geometry between the two views. The lines $p' - \tilde{p}'$ are all parallel to each other (can be easily seen from Figure 2.1) and are known as the epipolar lines. The epipolar lines are parallel because we are assuming parallel projection — in the general case epipolar lines converge to a point known as the epipole which is at the intersection of the line connecting the two centers of projection and the image plane (to be discussed in more detail in the next chapter).

The epipolar geometry, i.e., the direction of epipolar lines in both image planes, contains all the information regarding the viewing transformation between the two camera locations. We can see that in orthographic projection this information is generally not sufficient for uniquely determining the rigid motion of the camera. This is because the component of rotation around the axis perpendicular to the direction of the epipolar lines cannot be determined by the epipolar transformation (the transformation that maps epipolar lines in one image onto the epipolar lines of the other image). The epipolar geometry is, therefore, weaker than the full viewing transformation. In other words, the alignment transformation that is responsible for re-projecting the model onto a novel view, does not require a full recovery of the viewing transformation.

Corollary 1 *Recognition via alignment using two parallel projected model views can be achieved by recovering affine coordinates or by recovering affine structure and the epipolar geometry between the model and the novel view.*

In some cases epipolar geometry alone is sufficient for re-projecting the model onto the novel view. This can be done by intersecting the epipolar lines between each of the model views and the novel view. This is possible as long as the center of projection of the novel camera position is not collinear with the centers of projection of the two model views, or, in other words, the projection of the two axes of rotation of the viewing transformations between the model views and the novel view do not coincide. Re-projection via epipolar intersection may also be unstable for viewing transformations that are nearly singular (intersecting lines are nearly parallel).

The methods described so far were based on recovering some from of relative structure and/or the epipolar geometry resulting from the viewing transformations between the camera positions. The method discussed in the next section, the linear combination of views, achieves re-projection by employing a direct connection between the views of the same object.

2.5 The Linear Combination of Views and Affine Structure

Ullman and Basri (1989), (also Poggio 1990) have discovered a simple linear relationship between the corresponding points p, p' and p'' of three distinct views of the point P, under parallel projection.

Let o, o', o'', the projections of an arbitrary point O, be the origins of the three image coordinate frames, and let $(x, y), (x', y'), (x'', y'')$ be the coordinates of the image vectors $op, o'p', o''p''$, respectively. Ullman and Basri show the following result:

$$x'' = \alpha_1 x + \alpha_2 y + \alpha_3 x'$$
$$y'' = \beta_1 x + \beta_2 y + \beta_3 x'$$

where the coefficients α_i, β_i are independent of the positions of p, p', p''. Similarly,

$$x'' = \gamma_1 x + \gamma_2 y + \gamma_3 y'$$
$$y'' = \delta_1 x + \delta_2 y + \delta_3 y'$$

where the coefficients γ_i, δ_i are independent of the positions of p, p', p''.

The linear combination result requires three more corresponding points, making altogether four points projected from non-coplanar points in space, to perform re-projection. Re-projection, therefore does not directly implicate, object structure or epipolar geometry. The linear combination method has also a practical advantage of allowing more than four points to solve for the coefficients in a least squares fashion. With affine structure, or affine coordinates, there is no easy way of providing a least squares solution, unless we assume that the additional corresponding points are coplanar with the reference plane. We show next that the linear combination of views result can be derived from affine structure and epipolar geometry.

The derivation of affine structure in the previous section concluded with,

$$o'p' = A(op) + \gamma_p \boldsymbol{w},$$

where $\boldsymbol{w} = o'p'_3 - o'\tilde{p}'_3$ is the epipolar line direction, and γ_p is invariant under viewing transformations. Similarly, let $o''p'' = B(op) + \gamma_p \boldsymbol{s}$ be derived from the first model view and the novel view using the same four reference points. We derive the following result:

Proposition 1 *The coefficients of the linear combination result are expressed by the following equations:*

$$\begin{array}{lll}
\alpha_1 = b_{11} - \alpha_3 a_{11} & \alpha_2 = b_{12} - \alpha_3 a_{12} & \alpha_3 = \frac{w_1}{s_1} \\
\beta_1 = b_{21} - \beta_3 a_{11} & \beta_2 = b_{22} - \beta_3 a_{12} & \beta_3 = \frac{w_2}{s_1} \\
\gamma_1 = b_{11} - \gamma_3 a_{21} & \gamma_2 = b_{12} - \gamma_3 a_{22} & \gamma_3 = \frac{w_1}{s_2} \\
\delta_1 = b_{21} - \delta_3 a_{21} & \delta_2 = b_{22} - \delta_3 a_{22} & \delta_3 = \frac{w_2}{s_2}
\end{array}$$

Proof: Since γ_p is invariant and appears in both equations, we can cancel it and obtain the following equation:

$$\frac{1}{s_1}(x' - a_{11}x - a_{12}y) = \frac{1}{s_2}(y' - a_{21}x - a_{22}y) = \frac{1}{w_1}(x'' - b_{11}x - b_{12}y) = \frac{1}{w_2}(y'' - b_{21}x - b_{22}y)$$

$\square$

We therefore see that x'' and y'' can be represented as a linear combination of x, y, x' provided that the rotation between the model views has a non-zero component around the horizontal axis ($s_1 \neq 0$), and similarly that x'' and y'' can be represented as a linear combination of x, y, y' provided that the rotation between the model views has a non-zero component around the vertical axis ($s_2 \neq 0$). We also see that there is no restriction on the viewing transformation between the model views and the novel view ($\boldsymbol{w} = 0$ corresponds to the case of pure rotation around the optical axis, in which case the novel view is simply a 2D affine transformation of the model views).

2.6 Discussion

We described the geometrical concepts underlying non-metric structure, epipolar geometry, and alignment from two model views in the case of parallel projection. We wish to emphasize the following points.

The first point is the question of structure representation. We have seen two different, but mathematically equivalent, ways of representing structure from two parallel views: one is by affine coordinates and the other is by a geometric invariant. The geometric invariant approach led to a simpler method of re-projection; however, this is not the main point. The structure-by-coordinates approach necessarily involves the center of projection, which in the case of parallel projection can be any point on the object. Therefore, the two representations measure shape relative to a basis imposed on the object. In the general case of central projection, however, the structure-by-coordinates approach would result in shape relative to the camera coordinate frame, whereas the structure-by-geometric-invariant approach does not. The question of how to avoid implicating the center of projection is not only important for reasons of stability of reconstruction and recognition, but also, as described in the next chapter, is the key for allowing parallel and central projection to coexist in a single unified framework. It is important, therefore, to make a distinction between these two approaches.

The second point we wish to emphasize is the distinction between recovering the viewing transformation (the parameters of camera motion) and recovering the epipolar geometry. In parallel projection, the question of recovering the viewing transformation does not arise because two views are not sufficient for a unique recovery. In the more general case, however, we have a choice between recovering the viewing transformation or simply recovering the epipolar geometry. We saw that epipolar geometry is sufficient for recovering non-metric structure and for achieving recognition. This result extends, as described in the next chapter, to the general case of central projection.

Finally, we emphasize the connection between affine structure and epipolar geometry and the linear combination result of Ullman and Basri. The linear combination method of re-projection is by far the most efficient. Not only are structure and epipolar geometry not directly involved in the process, but in addition, more than the minimal necessary number of points can be used to recover a least squares solution to the re-projection problem. However, it may be difficult to extend this result to central projection without first directly addressing the issues of structure and camera geometry. The importance of Proposition 1 is to show that a more direct connection between distinct views of the same object (i.e., the linear combination result) can be derived in terms of structure and epipolar geometry. This suggests, therefore, that a similar connection may be derived in the general case of central projection if we can first represent the relation between corresponding points in a way that involves a non-metric invariant and epipolar geometry.

Projective Structure and Alignment in the General Case of Central Projection

Chapter 3

In this chapter we continue to pursue the geometric relation between objects and their views by considering the general problem of central projection. First, we would like to recover some form of non-metric structure from two views produced by central projection, and secondly we would like to achieve recognition from two model views and a novel view.

The problem may be approached in several ways. Therefore, before we attend to the proposed solution we will consider several questions. The first question is whether it is really necessary to work in a non-metric framework. In other words, what are the major problems in previous metric approaches? The second question concerns the requirements of a good representation of non-metric structure. In parallel projection, the two representations (coordinates and geometric invariants) were mathematically equivalent, but in the general case they may not be. Thirdly, what are the challenges in going from the relatively simple domain of parallel projection to central projection? For example, we will see that the affine structure invariant, defined with respect to a reference plane and a reference point, critically relies on the projection being parallel and does not apply to central projection. Furthermore, there is the issue that the transformation due to a plane under central projection requires four coplanar points, rather than three. Since four arbitrarily chosen points are generally not coplanar, the question is whether extending the geometric invariant approach is worthwhile from a practical point of view.

The solution we propose is based on a geometric invariant approach, rather than on recovering projective coordinates of the scene. We show that it is first necessary to define another kind of geometric invariant, different from the one proposed by Koenderink and Van Doorn (1991). The new structure invariant applies to both parallel and central projections, and similarly to parallel projection, requires only four non-coplanar points for its definition. The difference between the two cases (parallel and central projections) is that in central projection we must first recover the epipolar geometry between the two views. Once the location of epipoles is recovered there is essentially no difference between the two cases.

Our proposed solution has several advantages. First and foremost, we treat parallel and central projection alike, i.e., no distinction is made between the two at the algorithmic level. Secondly, there is no need for assuming internal camera calibration. On the other hand, if the cameras are calibrated, then rigidity of the object can be verified in the course of computing structure and during recognition. Thirdly, the computations are simple and linear.

3.1 Problems with Metric Approaches

The derivation of affine representations of structure, coordinates or geometric invariants, for purposes of SFM and recognition has a clear practical aspect when it comes to parallel projection: non-metric SFM can be achieved from two views (Koenderink & Van Doorn, 1991), rather than three views required for recovering metric structure, and recognition using the alignment approach can be achieved from two model images, rather than three (Ullman & Basri, 1989).

This advantage, of working with two rather than three views, is not present under perspective projection, however. It is known that two perspective views are sufficient for recovering metric structure (Roach & Aggarwal 1979, Longuett-Higgins 1981, Tsai & Huang 1984, Faugeras & Maybank 1990, Horn 1990, Horn 1991). The question, therefore, is why look for alternative representations of structure?

There are three major problems in structure from motion methods: (i) critical dependence on an orthographic or perspective model of projection, (ii) internal camera calibration, and (iii) the problem of stereo-triangulation.

The first problem is the strict division between methods that assume orthographic projection and methods that assume perspective projection. These two classes of methods do not overlap in their domain of application. The perspective model operates under conditions of significant perspective distortions, such as driving on a stretch of highway, requires a relatively large field of view and relatively large depth variations between scene points (Adiv 1989, Dutta & Synder 1990, Tomasi 1991, Broida *et al.* 1990). The orthographic model, on the other hand, provides a reasonable approximation when the imaging situation is at the other extreme, i.e., small field of view and small depth variation between object points (a situation for which perspective schemes often break down). Typical imaging situations are at neither end of these extremes and, therefore, would be vulnerable to errors in both models. From the standpoint of performing recognition, this problem implies that the viewer has control over his field of view — a property that may be reasonable to assume at the time of model acquisition, but less reasonable to assume occurring at recognition time.

The second problem is related to internal camera calibration. As discussed in Section 1.6, with perspective projection we assume a known and fixed relationship, described by five internal parameters, between the image plane and the camera coordinate system. In practice, these internal parameters must be calibrated prior to making use of the perspective model for recovering shape or for purposes of recognition. Although the calibration process is somewhat tedious, it is sometimes necessary for many of the available commercial cameras (Brown 1971, Faig 1975, Lenz and Tsai 1987, Faugeras, Luong and Maybank 1992).

The third problem is related to the way shape is typically represented under the perspective projection model. Because the center of projection is also the origin of the coordinate system for describing shape, the shape difference (e.g., difference in depth, between two object points), is orders of magnitude smaller than the distance to the scene, and this makes the computations very sensitive to noise. The sensitivity to noise is reduced if images are taken from distant viewpoints (large base-line in stereo triangulation), but that makes the process of establishing correspondence between points in both views more of a problem, and hence, may make the situation even worse. This problem does not occur under the assumption of orthographic projection because translation in depth is lost under orthographic projection, and therefore, the origin of the coordinate system for describing shape (metric and non-metric) is object-centered, rather than viewer-centered (Tomasi, 1991).

These three problems form the basis of evaluating the possible ways one can extend from parallel projection to central projection. It is clear that by replacing the perspective projection model with central projection we no longer have the calibration problem to be concerned with. The other two problems imply that we should avoid implicating the camera's center of projection with the definition of structure. We describe these issues in more detail in the next section.

3.2 From parallel to Central projection: Points of Interest

We saw in the previous chapter that in parallel projection the two representations of structure, i.e., coordinates and geometric invariants, were mathematically equivalent. The equivalence comes from the unique property that depth translation is lost under parallel projection. Therefore, structure is measured relative to an object-centered frame of reference, rather than relative to a camera-centered frame. Since this property does not hold under central (and perspective) projection, then representing structure by projective coordinates relies on a camera-centered frame of reference, whereas a geometric-invariant approach may still provide us with an object-centered representation. The stereo-triangulation problem, for example, clearly favors the latter approach.

Recent work in this area approach the problem of recovering non-metric structure from the standpoint of recovering projective coordinates of the object in 3D projective space (Faugeras 1992, Mohr, Quan, Veillon & Boufama 1992, Hartley, Gupta & Chang 1992). The general idea is motivated by the result that five points in general position provide a basis in 3D projective space (see, for instance, Semple & Kneebone 1952). Faugeras, for example, shows that given the epipoles and five corresponding points between the two views one can recover the projective homogeneous coordinates of all other points projecting to corresponding points in both images. Along with recovering projective coordinates, Faugeras recovers the center of projection and the full set of 11 parameters describing the camera geometry. This approach of representing structure by projective coordinates clearly provides a solution to the internal calibration problem, but not for the other two problems. Parallel projection, for instance, is a point of singularity for these methods.

We take on a different path and represent structure by defining a geometric invariant that applies to both parallel and central projections. We wish to extend, therefore, the construction of affine structure to handle also the case of central projection. The are generally two problems in taking this path. The first problem is that affine structure was defined in a way that critically relies on the properties of parallel projection. Consider, for example, Figure 3.1 which illustrates the same geometric construction as in Figure 2.1 but under central projection. The relationship between the points $P, P_3, \tilde{P}, \tilde{P}_3$ and the center of projection V_1 and the points $p', \tilde{p}', p'_3, \tilde{p}'_3, V_l$, where V_l denotes the epipole, can be described as a perspectivity between two triangles. We have five points in each triangle which would seem as sufficient for defining an invariant relation (since four points determine the projectivity of the plane in projective geometry, then the coordinates of the fifth point are invariant). This, however, requires that no three of the points be collinear — a requirement that is not satisfied in this case.

Secondly, the idea of using a reference plane is problematic. In affine geometry three points are sufficient for uniquely determining the correspondences of all other points on the reference plane. In projective geometry, however, four coplanar points are required. Since

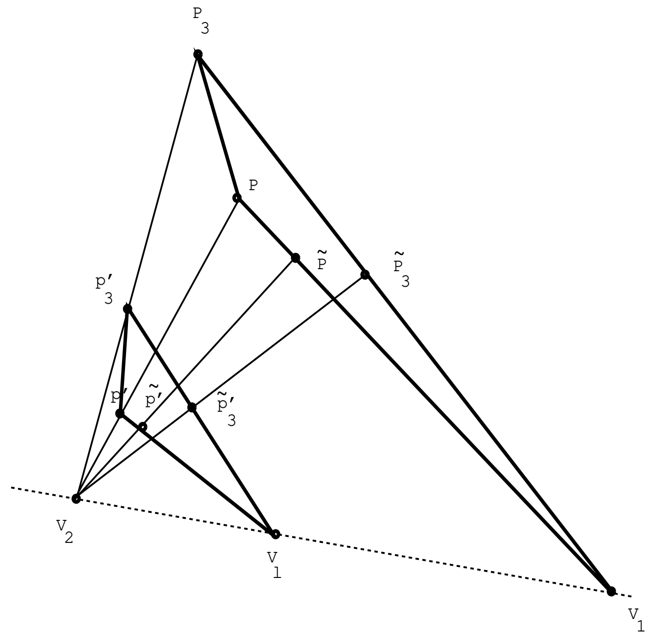


Figure 3.1: Affine structure invariant does not hold under central projection (see text).

four arbitrarily chosen points are generally not coplanar, this raises a question of whether extending the structure-by-geometric-invariant approach is worthwhile from a practical point of view.

We first describe a different affine structure invariant under parallel projection using two reference planes, rather than one reference plane and a reference point. The new affine invariant applies to central projection as well. We then show that, given the epipoles, only three corresponding points for each reference plane are sufficient for recovering the associated projective transformations induced by those planes. This leads to the main result (Theorem 1) that, in addition to the epipoles, only four corresponding points, projected from four non-coplanar points in the scene, are sufficient for recovering the projective structure invariant for all other points.

3.3 Affine Structure Using Two Reference Planes

We make use of the same information — the projections of four non-coplanar points — to set up two reference planes. Let P_j, $j = 1, ..., 4$, be the four non-coplanar reference points in space, and let $p_j \longleftrightarrow p'_j$ be their observed projections in both views. The points P_1, P_2, P_3 and P_2, P_3, P_4 lie on two different planes, therefore, we can account for the motion of all

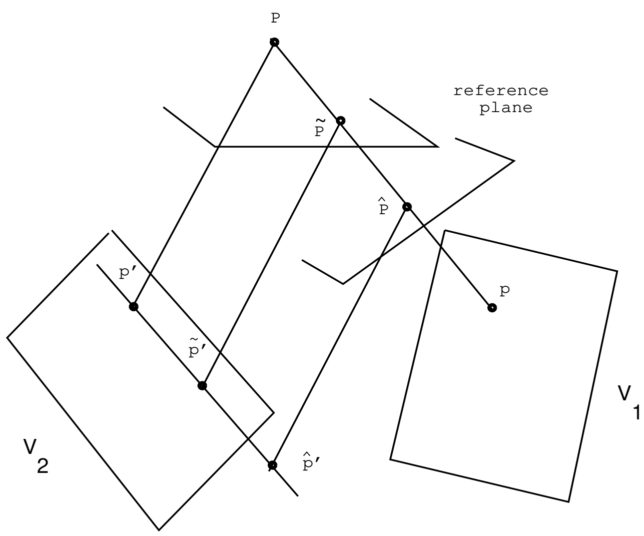


Figure 3.2: Affine structure using two reference planes.

points coplanar with each of these two planes. Let P be a point of interest, not coplanar with either of the reference planes, and let $\tilde{P}$ and $\hat{P}$ be its projections onto the two reference planes along the ray towards the first view.

Consider Figure 3.2. The projection of $P, \tilde{P}$ and $\hat{P}$ onto $p', \tilde{p}'$ and $\hat{p}'$ respectively, gives rise to two similar trapezoids from which we derive the following relation:

$$\alpha_p = \frac{|P - \tilde{P}|}{|P - \hat{P}|} = \frac{|p' - \tilde{p}'|}{|p' - \hat{p}'|}.$$

The ratio α_p is invariant under parallel projection. Similar to the case of Koenderink and Van Doorn's affine structure γ_p, we can easily show a one-to-one mapping between the affine coordinates of P and α_p:

$$b_3 = \frac{\alpha_p}{1 - \alpha_p} \frac{\hat{p}'_4 - \tilde{p}'_4}{\hat{p}' - \tilde{p}'}.$$

There is no particular advantage for preferring α_p over γ_p as a measure of affine structure, but as will be described below, this new construction forms the basis for extending affine structure to projective structure: together with the epipole, the similarity of trapezoids in the affine case turns into a cross-ratio in the projective case.

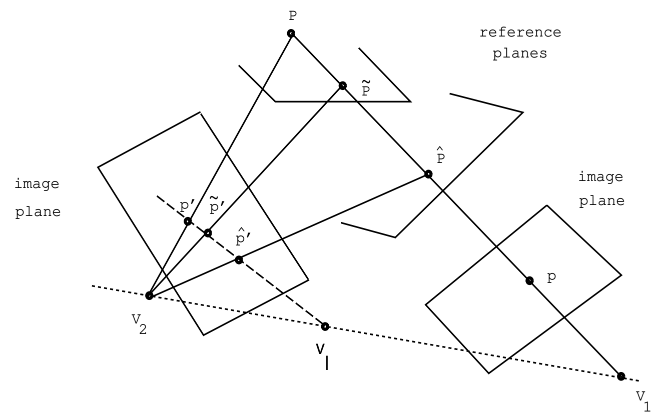


Figure 3.3: Definition of projective shape as the cross ratio of $p', \tilde{p}', \hat{p}', V_l$.

3.4 Projective Structure

We assume for now that the location of both epipoles is known, and we will address the problem of finding the epipoles later. The epipoles, also known as the foci of expansion, are the intersections of the line in space connecting the two centers of projection and the image planes. There are two epipoles, one on each image plane — the epipole on the second image we call the left epipole, and the epipole on the first image we call the right epipole. The image lines emanating from the epipoles are known as the epipolar lines.

Consider Figure 3.3 which illustrates the two reference plane construction, defined earlier for parallel projection, now displayed in the case of central projection. The left epipole is denoted by V_l, and because it is on the line V_1V_2 (connecting the two centers of projection), the line PV_1 projects onto the epipolar line $p'V_l$. Therefore, the points $\tilde{P}$ and $\hat{P}$ project onto the points $\tilde{p}'$ and $\hat{p}'$, which are both on the epipolar line $p'V_l$. The points $p', \tilde{p}', \hat{p}'$ and V_l are collinear and projectively related to $P, \tilde{P}, \hat{P}, V_1$, and therefore have the same cross-ratio:

$$\alpha_p = \frac{|P - \tilde{P}|}{|P - \hat{p}|} \cdot \frac{|V_1 - \hat{p}|}{|V_1 - \tilde{P}|} = \frac{|p' - \tilde{p}'|}{|p' - \hat{p}'|} \cdot \frac{|V_l - \hat{p}'|}{|V_l - \tilde{p}'|}.$$

Note that when the epipole V_l becomes an ideal point (vanishes along the epipolar line), then α_p is the same as the affine invariant defined in section 3.3 for parallel projection.

The cross-ratio α_p is a direct extension of the affine structure invariant defined in section 3.3 and is referred to as *projective structure*. We can use this invariant to reconstruct any novel view of the object (taken by a non-rigid camera) without ever recovering depth or even projective coordinates of the object.

Having defined the projective shape invariant, and assuming we still are given the locations of the epipoles, we show next how to recover the projections of the two reference planes onto the second image plane, i.e., we describe the computations leading to $\tilde{p}'$ and $\hat{p}'$.

Since we are working under central projection, we need to identify four coplanar points on each reference plane. In other words, in the projective geometry of the plane, four corresponding points, no three of which are collinear, are sufficient to determine uniquely all other correspondences (see Appendix A, for more details). We must, therefore, identify four corresponding points that are projected from four coplanar points in space, and then recover the projective transformation that accounts for all other correspondences induced from that plane. The following proposition states that the corresponding epipoles can be used as a fourth corresponding point for any three corresponding points selected from the pair of images.

Proposition 2 *A projective transformation, A, that is determined from three arbitrary, non-collinear, corresponding points and the corresponding epipoles, is a projective transformation of the plane passing through the three object points which project onto the corresponding image points. The transformation A is an induced epipolar transformation, i.e., the ray Ap intersects the epipolar line $p'V_l$ for any arbitrary image point p and its corresponding point p'.*

Comment: An epipolar transformation F is a mapping between corresponding epipolar lines and is determined (not uniquely) from three corresponding epipolar lines and the epipoles. The induced point transformation is $E = (F^{-1})^t$ (induced from the point/line duality of projective geometry, see Appendix C for more details on epipolar transformations).

Proof: Let $p_j \longleftrightarrow p'_j$, $j = 1, 2, 3$, be three arbitrary corresponding points, and let V_l and V_r denote the left and right epipoles. First note that the four points p_j and V_r are projected from four coplanar points in the scene. The reason is that the plane defined by the three object points P_j intersects the line V_1V_2 connecting the two centers of projection, at a point — regular or ideal. That point projects onto both epipoles. The transformation A, therefore, is a projective transformation of the plane passing through the three object points P_1, P_2, P_3. Note that A is uniquely determined provided that no three of the four points are collinear.

Let $\mu\tilde{p}' = Ap$ for some arbitrary point p. Because lines are projective invariants, any point along the epipolar line pV_r must project onto the epipolar line $p'V_l$. Hence, A is an induced epipolar transformation. $\square$

Given the epipoles, therefore, we need just three points to determine the correspondences of all other points coplanar with the reference plane passing through the three corresponding object points. The transformation (collineation) A is determined from the following equations:

$$Ap_j = \rho_j p'_j, \qquad j = 1,2,3$$
$$AV_r = \rho V_l,$$

where ρ, ρ_j are unknown scalars, and $A_{3,3} = 1$. One can eliminate ρ, ρ_j from the equations and solve for the matrix A from the three corresponding points and the corresponding epipoles. That leads to a linear system of eight equations, and is described in more detail in Appendix A.

If P_1, P_2, P_3 define the first reference plane, the transformation A determines the location of $\tilde{p}'$ for all other points p ($\tilde{p}'$ and p' coincide if P is coplanar with the first reference plane). In other words, we have that $\tilde{p}' = Ap$. Note that $\tilde{p}'$ is not necessarily a point on the second image plane, but it is on the line $V_2\tilde{P}$. We can determine its location on the second plane by normalizing Ap such that its third component is set to 1.

Similarly, let P_2, P_3, P_4 define the second reference plane (assuming the four object points P_j, $j = 1, ..., 4$, are non-coplanar). The transformation E is uniquely determined by the equations

$$Ep_j = \rho_j p'_j, \qquad j = 2,3,4$$
$$EV_r = \rho V_l,$$

and determines all other correspondences induced by the second reference plane (we assume that no three of the four points used to determine E are collinear). In other words, Ep determines the location of $\hat{p}'$ up to a scale factor along the ray $V_2\hat{P}$.

Instead of normalizing Ap and Ep we compute α_p from the cross-ratio of the points represented in homogeneous coordinates, i.e., the cross-ratio of the four rays $V_2p', V_2\tilde{p}', V_2\hat{p}', V_2V_l$, as follows: Let the rays p', V_l be represented as a linear combination of the rays $\tilde{p}' = Ap$ and $\hat{p}' = Ep$, i.e.,

$$p' = \tilde{p}' + k\hat{p}'$$
$$V_l = \tilde{p}' + k'\hat{p}',$$

then $\alpha_p = \frac{k}{k'}$ (see Appendix B for more details). This way of computing the cross-ratio is preferred over the more familiar cross-ratio of four collinear points, because it enables us to work with all elements of the projective plane, including ideal points (a situation that arises, for instance, when epipolar lines are parallel, and in general under parallel projection).

We have therefore shown the following result:

Theorem 1 *In the case where the location of epipoles are known, then four corresponding points, coming from four non-coplanar points in space, are sufficient for computing the projective structure invariant α_p for all other points in space projecting onto corresponding points in both views, for all central projections, including parallel projection.*

This result shows that the difference between parallel and central projection lies entirely on the epipoles. In both cases four non-coplanar points are sufficient for obtaining the invariant, but in the parallel projection case we have prior knowledge that both epipoles are ideal, therefore they are not required for determining the transformations A and E (in other words, A and E are affine transformations, more on that in Section 3.7). Also, because cross-ratios are invariant under projectivities it is clear why we included in our non-rigid camera model (Section 1.6) the capability to take projections of projections (α_p remains unchanged under any projective transformation of the left image).

We next discuss algorithms for recovering the location of epipoles. The problem of recovering the epipoles is well known and several approaches have been suggested in the past (Longuet-Higgins and Prazdny 1980, Rieger-Lawton 1985, Faugeras and Maybank 1990, Hildreth 1991, Horn 1990, Faugeras 1992, Faugeras, Luong and Maybank 1992).

In general, the epipoles can be recovered from six points (four of which are assumed to be coplanar), seven points (non-linear algorithm, see Faugeras & Maybank 1990), or, as recently shown by Faugeras (1992), eight points. We start with the six-point method (two additional points to the four we already have). The method is a direct extension of the Koenderink and Van Doorn (1991) construction in parallel projection, and was described earlier by Lee (1988) for the purpose of recovering the translational component of camera motion. The second algorithm we describe for locating the epipoles requires eight points and is based on the fundamental matrix of Longuet-Higgins (1981).

3.5 Epipoles from Six Points

We can recover the correspondences induced from the first reference plane by selecting four corresponding points, assuming they are projected from four coplanar object points. Let

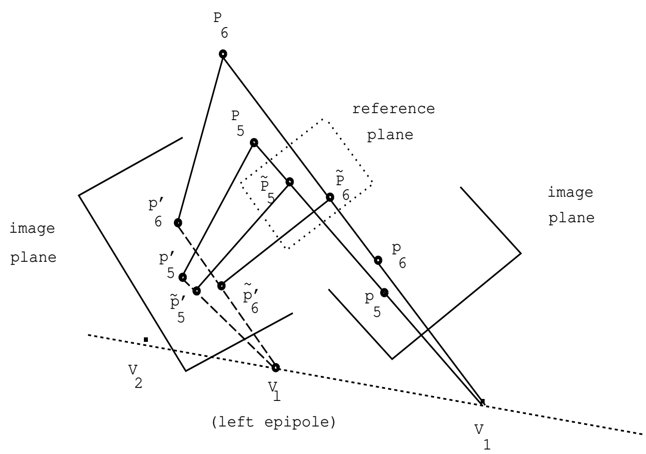


Figure 3.4: The geometry of locating the left epipole using two points out of the reference plane.

$p_j = (x_j, y_j, 1)$ and $p'_j = (x'_j, y'_j, 1)$ and $j = 1, ..., 4$ represent the standard image coordinates of the four corresponding points, no three of which are collinear, in both projections. Therefore, the transformation A is uniquely determined by the following equations,

$$\rho_j p'_j = A p_j.$$

Let $\tilde{p}' = Ap$ be the homogeneous coordinate representation of the ray $V_2\tilde{P}$, and let $\tilde{p}^{-1} = A^{-1}p'$.

Having accounted for the motion of the reference plane, we can easily find the location of the *epipoles* (in standard coordinates). Given two object points P_5, P_6 that are *not* on the reference plane, we can find both epipoles by observing that $\tilde{p}'$ is on the left epipolar line, and similarly that $\tilde{p}^{-1}$ is on the right epipolar line. Stated formally, we have the following proposition:

Proposition 3 *The left epipole, denoted by V_l, is at the intersection of the line $p'_5\tilde{p}'_5$ and the line $p'_6\tilde{p}'_6$. Similarly, the right epipole, denoted by V_r, is at the intersection of $p_5\tilde{p}_5^{-1}$ and $p_6\tilde{p}_6^{-1}$.*

Proof: It is sufficient to prove the claim for one of the epipoles, say the left epipole. Consider Figure 3.4 which describes the construction geometrically. By construction, the line $P_5\tilde{P}_5V_1$ projects to the line $p'_5\tilde{p}'_5$ via V_2 (points and lines are projective invariants) and therefore they are coplanar. In particular, V_1 projects to V_l which is located at the intersection of $p'_5\tilde{p}'_5$ and V_1V_2. Similarly, the line $p'_6\tilde{p}'_6$ intersects V_1V_2 at $\hat{V}_l$. Finally, V_l and $\hat{V}_l$ must coincide because the two lines $p'_5\tilde{p}'_5$ and $p'_6\tilde{p}'_6$ are coplanar (both are on the image plane). $\square$

Algebraically, we can recover the ray V_1V_2, or V_l up to a scale factor, using the following formula:

$$V_l = (p'_5 \times \tilde{p}'_5) \times (p'_6 \times \tilde{p}'_6).$$

Note that V_l is defined with respect to the standard coordinate frame of the second camera. We treat the epipole V_l as the ray V_1V_2 with respect to V_2, and the epipole V_r as the same ray but with respect to V_1. Note also that the third component of V_l is zero if epipolar lines are parallel, i.e., V_l is an ideal point in projective terms (happening under parallel projection, or when the non-rigid camera motion brings the image plane to a position where it is parallel to the line V_1V_2).

In the case where more than two epipolar lines are available (such as when more than six corresponding points are available), one can find a least-squares solution for the epipole by using a principle component analysis, as follows. Let B be a $k \times 3$ matrix, where each row represents an epipolar line. The least squares solution to V_l is the unit eigenvector associated with the smallest eigenumber of the 3×3 matrix B^tB. Note that this can be done analytically because the characteristic equation is a cubic polynomial.

Altogether, we have a six point algorithm for recovering both the epipoles, and the projective structure α_p, and for performing re-projection onto any novel view. We summarize in the following section the 6-point algorithm.

3.5.1 Re-projection Using Projective Structure: 6-point Algorithm

We assume we are given two model views of a 3D object, and that all points of interest are in correspondence. We assume these correspondences can be based on measures of correlation, as used in optical-flow methods (see also Chapter 7 for methods for extracting correspondences using combination of optical flow and affine geometry).

Given a novel view we extract six corresponding points (with one of the model views): $p_j \longleftrightarrow p'_j \longleftrightarrow p''_j$, $j = 1, ..., 6$. We assume the first four points are projected from four coplanar points, and the other corresponding points are projected from points that are not on the reference plane. Without loss of generality, we assume the standard coordinate representation of the image planes, i.e., the image coordinates are embedded in a 3D vector whose third component is set to 1 (see Appendix A). The computations for recovering projective shape and performing re-projection are described below.

1: Recover the transformation A that satisfies $\rho_j p'_j = Ap_j$, $j = 1, ..., 4$. This requires setting up a linear system of eight equations (see Appendix A). Apply the transformation to all points p, denoting $\tilde{p}' = Ap$. Also recover the epipoles $V_l = (p'_5 \times \tilde{p}'_5) \times (p'_6 \times \tilde{p}'_6)$ and $V_r = (p_5 \times A^{-1}p'_5) \times (p_6 \times A^{-1}p'_6)$.

2: Recover the transformation E that satisfies $\rho V_l = EV_r$ and $\rho_j p'_j = Ep_j$, $j = 4, 5, 6$.

3: Compute the cross-ratio of the points p', Ap, Ep, V_l, for all points p and denote that by α_p (see Appendix B for details on computing the cross-ratio of four rays).

4: Perform step 1 between the first and novel view: recover $\tilde{A}$ that satisfies $\rho_j p''_j = \tilde{A}p_j$, $j = 1, ..., 4$, apply $\tilde{A}$ to all points p and denote that by $\tilde{p}'' = \tilde{A}p$, recover the epipoles $V_{ln} = (p''_5 \times \tilde{p}''_5) \times (p''_6 \times \tilde{p}''_6)$ and $V_{rn} = (p_5 \times \tilde{A}^{-1}p''_5) \times (p_6 \times \tilde{A}^{-1}p''_6)$.

5: Perform step 2 between the first and novel view: Recover the transformation $\tilde{E}$ that satisfies $\rho V_{ln} = \tilde{E}V_{rn}$ and $\rho_j p''_j = Ep_j$, $j = 4, 5, 6$.

6: For every point p, recover p'' from the cross-ratio α_p and the three rays $\tilde{A}p, \tilde{E}p, V_{ln}$. Normalize p'' such that its third coordinate is set to 1.

The entire procedure requires setting up a linear system of eight equations four times (Step 1,2,4,5) and computing cross-ratios (linear operations as well).

The results so far required prior knowledge (or assumption) that four of the corresponding points are coming from coplanar points in space. This requirement can be avoided, using two more corresponding points (making eight points overall), and is described in the next section.

3.6 Epipoles from Eight Points

We adopt a recent algorithm suggested by Faugeras (1992) which is based on Longuet-Higgins' (1981) fundamental matrix. The method is very simple and requires eight corresponding points for recovering the epipoles.

Let F be an epipolar transformation, i.e., $Fl = \mu l'$, where $l = V_r \times p$ and $l' = V_l \times p'$ are corresponding epipolar lines. We can rewrite the projective relation of epipolar lines using the matrix form of cross-products:

$$F(V_r \times p) = F[V_r]p = \rho l',$$

where $[V_r]$ is a skew symmetric matrix (and hence has rank 2). From the point/line incidence property we have that $p' \cdot l' = 0$ and therefore, $p'^t F[V_r]p = 0$, or $p'^t Hp = 0$ where $H = F[V_r]$. The matrix H is known as the fundamental matrix introduced by Longuet-Higgins (1981), and is of rank 2. One can recover H (up to a scale factor) directly from eight corresponding points, or by using a principle components approach if more than eight points are available. Finally, it is easy to see that

$$HV_r = 0,$$

and therefore the epipole V_r can be uniquely recovered (up to a scale factor). Note that the determinant of the first principle minor of H vanishes in the case where V_r is an ideal point, i.e., $h_{11}h_{22} - h_{12}h_{21} = 0$. In that case, the x, y components of V_r can be recovered (up to a scale factor) from the third row of H. The epipoles, therefore, can be uniquely recovered under both central and parallel projection. We have arrived at the following theorem:

Theorem 2 *In the case where we have eight corresponding points of two views taken under central projection (including parallel projection), four of these points, coming from four non-coplanar points in space, are sufficient for computing the projective structure invariant α_p for the remaining four points and for all other points in space projecting onto corresponding points in both views.*

We summarize in the following section the 8-point scheme for reconstructing projective structure and performing re-projection onto a novel view.

3.6.1 8-point Re-projection Algorithm

We assume we have eight corresponding points between two model views and the novel view, $p_j \longleftrightarrow p'_j \longleftrightarrow p''_j$, $j = 1, ..., 8$, and that the first four points are coming from four non-coplanar points in space. The computations for recovering projective structure and performing re-projection are described below.

1: Recover the fundamental matrix H (up to a scale factor) that satisfies ${p'_j}^t Hp_j$, $j = 1, ..., 8$. The right epipole V_r then satisfies $HV_r = 0$. Similarly, the left epipole is recovered from the relation $p^t \tilde{H} p'$ and $\tilde{H} V_l = 0$.

2: Recover the transformation A that satisfies $\rho V_l = AV_r$ and $\rho_j p'_j = Ap_j$, $j = 1, 2, 3$. Similarly, recover the transformation E that satisfies $\rho V_l = EV_r$ and $\rho_j p'_j = Ep_j$, $j = 2, 3, 4$.

3: Compute α_p as the cross-ratio of p', Ap, Ep, V_l, for all points p.

4: Perform step 1 and 2 between the first and novel view: recover the epipoles V_{rn}, V_{ln}, and the transformations $\tilde{A}$ and $\tilde{E}$.

5: For every point p, recover p'' from the cross-ratio α_p and the three rays $\tilde{A}p, \tilde{E}p, V_{ln}$. Normalize p'' such that its third coordinate is set to 1.

We discuss below an important property of this procedure which is the transparency with respect to projection model: central and parallel projection are treated alike — a property which has implications on stability of re-projection no matter what degree of perspective distortions are present in the images.

3.7 The Case of Parallel Projection

The construction for obtaining projective structure is well defined for all central projections, including the case where the center of projection is an ideal point, i.e., such as happening with parallel projection. The construction has two components: the first component has to do with recovering the epipolar geometry via reference planes, and the second component is the projective invariant α_p.

From Proposition 2 the projective transformations A and E can be uniquely determined from three corresponding points and the corresponding epipoles. If both epipoles are ideal, the transformations become affine transformations of the plane (an affine transformation separates ideal points from Euclidean points). All other possibilities (both epipoles are Euclidean, one epipole Euclidean and the other epipole ideal) lead to projective transformations. Because a projectivity of the projective plane is uniquely determined from any four points on the projective plane (provided no three are collinear), the transformations A and E are uniquely determined under all situations of central projection — including parallel projection.

The projective invariant α_p is the same as the one defined under parallel projection (Section 3.3) — affine structure is a particular instance of projective structure in which the epipole V_l is an ideal point. By using the same invariant for both parallel and central projection, and because all other elements of the geometric construction hold for both projection models, the overall system is transparent to the projection model being used.

The first implication of this property has to do with stability. Projective structure does not require any perspective distortions, therefore all imaging situations can be handled — wide or narrow field of views. The second implication is that 3D visual recognition from 2D images can be achieved in a uniform manner with regard to the projection model. For instance, we can recognize (via re-projection) a perspective image of an object from only two orthographic model images, and in general any combination of perspective and orthographic images serving as model or novel views is allowed.

Similar to the case of parallel projection, we can achieve recognition just from the epipolar geometry, but under the condition that the centers of projection of the two model camera locations and the novel camera location are not collinear. This is a rather weak result as it implies instabilities in near singular situations. We describe this in more details in the following section.

3.8 On the Intersection of Epipolar Lines

Barret *et al.* (1991) derive a quadratic invariant based on Longuet-Higgins' fundamental matrix. We describe briefly their invariant and show that it is equivalent to performing re-projection using intersection of epipolar lines.

In section 3.6 we derived Longuet-Higgins' fundamental matrix relation $p'^t H p = 0$. Barret *et al.* note that the equation can be written in vector form $\boldsymbol{h}^t \cdot \boldsymbol{q} = 0$, where $\boldsymbol{h}$ contains the elements of H and

$$\boldsymbol{q} = (x'x, x'y, x', y'x, y'y, y', x, y, 1).$$

Therefore, the matrix

$$B = \begin{bmatrix} \boldsymbol{q}_1 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{q}_9 \end{bmatrix} \tag{3.1}$$

must have a vanishing determinant. Given eight corresponding points, the condition $|B| = 0$ leads to a constraint line in terms of the coordinates of any ninth point, i.e., $\alpha x + \beta y + \gamma =$

0. The location of the ninth point in any third view can, therefore, be determined by intersecting the constraint lines derived from views 1 and 3, and views 2 and 3.

Another way of deriving this re-projection method is by first noticing that H is a correlation that maps p onto the corresponding epipolar line $l' = V_l \times p'$ (see section 3.6). Therefore, from views 1 and 3 we have the relation

$$p''^t \tilde{H} p = 0,$$

and from views 2 and 3 we have the relation

$$p''^t \hat{H} p' = 0,$$

where $\tilde{H}p$ and $\hat{H}p'$ are two intersecting epipolar lines. Given eight corresponding points, we can recover $\tilde{H}$ and $\hat{H}$. The location of any ninth point p'' can be recovered by intersecting the lines $\tilde{H}p$ and $\hat{H}p'$.

This way of deriving the re-projection method has an advantage over using the condition $|B| = 0$ directly, because one can use more than eight points in a least squares solution (via SVD) for the matrices $\tilde{H}$ and $\hat{H}$.

Approaching the re-projection problem using intersection of epipolar lines is problematic for novel views that have a similar epipolar geometry to that of the two model views (these are situations where the two lines $\tilde{H}p$ and $\hat{H}p'$ are nearly parallel, such as when the object rotates around nearly the same axis for all views, or the centers of projection of the three cameras are nearly collinear). We therefore expect sensitivity to errors also under conditions of small separation between views. The method becomes more practical if one uses multiple model views instead of only two, because each model view adds one epipolar line and all lines should intersect at the location of the point of interest in the novel view.

We discuss next the possibility of working with a rigid camera (i.e., perspective projection and calibrated cameras).

3.9 The Rigid Camera Case

The advantage of the non-rigid camera model (or the central projection model) used so far is that images can be obtained from uncalibrated cameras. The price paid for this property is that the images that produce the same projective structure invariant (equivalence class of images of the object) can be produced by applying non-rigid transformations of the object, in addition to rigid transformations.

In this section we show that it is possible to verify whether the images were produced by rigid transformations, which is equivalent to working with perspective projection assuming the cameras are internally calibrated. This can be done for both schemes presented above, i.e., the 6-point and 8-point algorithms. In both cases we exclude orthographic projection and assume only perspective projection.

In the perspective case, the second reference plane is the image plane of the first model view, and the transformation for projecting the second reference plane onto any other view is the rotational component of camera motion (rigid transformation). We recover the rotational component of camera motion by adopting a result derived by Lee (1988), who shows that the rotational component of motion can be uniquely determined from two corresponding points and the corresponding epipoles. We then show that projective structure can be uniquely determined, up to a uniform scale factor, from two calibrated perspective images.

Proposition 4 (Lee, 1988) *In the case of perspective projection, the rotational component of camera motion can be uniquely recovered, up to a reflection, from two corresponding points and the corresponding epipoles. The reflection component can also be uniquely determined by using a third corresponding point.*

Proof: Let $l'_j = p'_j \times V_l$ and $l_j = p_j \times V_r$, $j = 1, 2$ be two corresponding epipolar lines. Because R is an orthogonal matrix, it leaves vector magnitudes unchanged, and we can normalize the length of l'_1, l'_2, V_l to be of the same length of l_1, l_2, V_r, respectively. We have therefore, $l'_j = Rl_j$, $j = 1, 2$, and $V_l = RV_r$, which is sufficient for determining R up to a reflection. Note that because R is a rigid transformation, it is both an epipolar and an induced epipolar transformation (the induced transformation E is determined by $E = (R^{-1})^t$, therefore $E = R$ because R is an orthogonal matrix).

To determine the reflection component, it is sufficient to observe a third corresponding point $p_3 \longleftrightarrow p'_3$. The object point P_3 is along the ray $V_1 p_3$ and therefore has the coordinates $\alpha_3 p_3$ (w.r.t. the first camera coordinate frame), and is also along the ray $V_2 p'_3$ and therefore has the coordinates $\alpha'_3 p'_3$ (w.r.t. the second camera coordinate frame). We note that the ratio between α_3 and α'_3 is a positive number. The change of coordinates is represented by:

$$\beta V_r + \alpha_3 R p_3 = \alpha'_3 p'_3,$$

where β is an unknown constant. If we multiply both sides of the equation by l'_j, $j = 1, 2, 3$, the term βV_r drops out, because V_r is incident to all left epipolar lines, and after

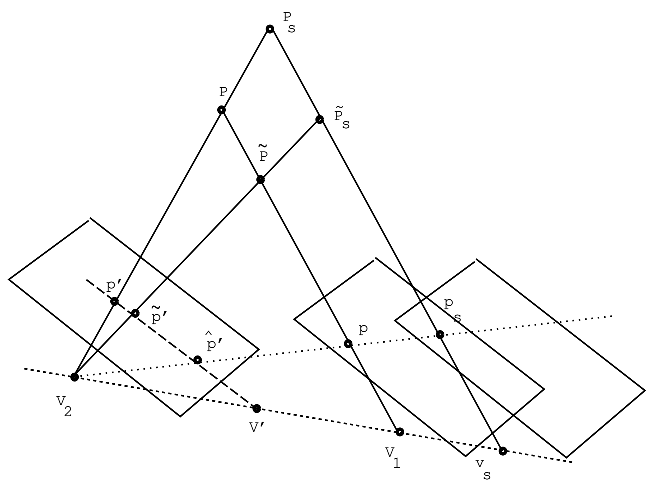


Figure 3.5: Illustration that projective shape can be recovered only up to a uniform scale (see text).

substituting l_j^t with ${l'_j}^t R$, we are left with,

$$\alpha_3 l_j^t \cdot p_3 = \alpha'_3 {l'_j}^t \cdot p'_3,$$

which is sufficient for determining the sign of l'_j. $\square$

The rotation matrix R can be uniquely recovered from any three corresponding points and the corresponding epipoles. Projective structure can be reconstructed by replacing the transformation E of the second reference plane, with the rigid transformation R (which is equivalent to treating the first image plane as a reference plane). We show next that this can lead to projective structure up to an unknown uniform scale factor (unlike the non-rigid camera case).

Proposition 5 *In the perspective case, the projective shape constant α_p can be determined, from two views, at most up to a uniform scale factor.*

Proof: Consider Figure 3.5, and let the effective translation be $V_2 - V_s = k(V_2 - V_1)$, which is the true translation scaled by an unknown factor k. Projective shape, α_p, remains fixed if the scene and the focal length of the first view are scaled by k: from similarity of triangles we have,

$$\begin{aligned} k = \frac{V_s - V_2}{V_1 - V_2} &= \frac{p_s - V_s}{p - V_1} = \frac{f_s}{1} \\ &= \frac{P_s - V_s}{P - V_1} = \frac{P_s - V_2}{P - V_2} \end{aligned}$$

where f_s is the scaled focal length of the first view. Since the magnitude of the translation along the line V_1V_2 is irrecoverable, we can assume it is null, and compute α_p as the cross-ratio of p', Ap, Rp, V_l which determines projective structure up to a uniform scale. $\square$

Because α_p is determined up to a uniform scale, we need an additional point in order to establish a common scale during the process of re-projection (we can use one of the existing six or eight points we already have). We obtain, therefore, the following result:

Theorem 3 *In the perspective case, a rigid re-projection from two model views onto a novel view is possible, using four corresponding points coming from four non-coplanar points, and the corresponding epipoles. The projective structure computed from two perspective images, is invariant up to an overall scale factor.*

Orthographic projection is excluded from this result because it is well known that the rotational component cannot be uniquely determined from two orthographic views (Ullman 1979, Huang and Lee 1989, Aloimonos and Brown 1989). To see what happens in the case of parallel projection note that the epipoles are vectors on the xy plane of their coordinate systems (ideal points), and the epipolar lines are two vectors perpendicular to the epipole vectors. The equation $RV_r = V_l$ takes care of the rotation in plane (around the optical axis). The other two equations $Rl_j = l'_j$, $j = 1, 2$, take care only of rotation around the epipolar direction — rotation around an axis perpendicular to the epipolar direction is not accounted for. The equations for solving for R provide a non-singular system of equations but do produce a rotation matrix with no rotational components around an axis perpendicular to the epipolar direction.

3.10 Simulation Results Using Synthetic Objects

We ran simulations using synthetic objects to illustrate the re-projection process using the 6-point scheme under various imaging situations. We also tested the robustness of the re-projection method under various types of noise. Because the 6-point scheme requires that four of the corresponding points be projected from four coplanar points in space, it is of special interest to see how the method behaves under conditions that violate this assumption, and under noise conditions in general. The stability of the 8-point algorithm largely depends on the method for recovering the epipoles. The method adopted from Faugeras (1992), described in Section 3.6, based on the fundamental matrix, tends to be

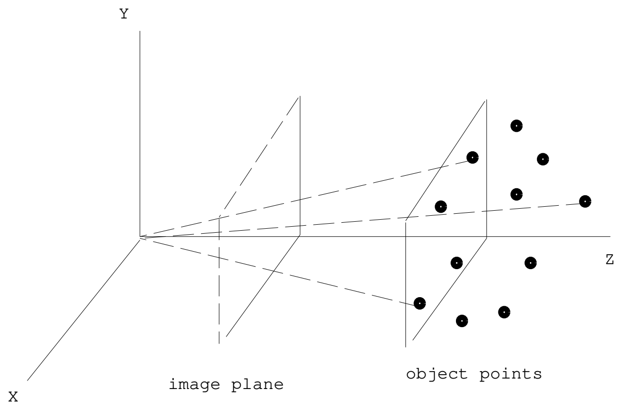


Figure 3.6: The basic object configuration for the experimental set-up.

very sensitive to noise if the minimal number of points (eight points) are used. We have, therefore, focused the experimental error analysis on the 6-point scheme.

Figure 3.6 illustrates the experimental set-up. The object consists of 26 points in space arranged in the following manner: 14 points are on a plane (reference plane) ortho-parallel to the image plane, and 12 points are out of the reference plane. The reference plane is located two focal lengths away from the center of projection (focal length is set to 50 units). The depth of out-of-plane points varies randomly between 10 to 25 units away from the reference plane. The x, y coordinates of all points, except the points $P_1, ..., P_6$, vary randomly between 0 — 240. The 'privileged' points $P_1, ..., P_6$ have x, y coordinates that place these points all around the object (clustering privileged points together will inevitably contribute to instability).

The first view is simply a perspective projection of the object. The second view is a result of rotating the object around the point $(128, 128, 100)$ with an axis of rotation described by the unit vector $(0.14, 0.7, 0.7)$ by an angle of 29 degrees, followed by a perspective projection (note that rotation about a point in space is equivalent to rotation about the center of projection followed by translation). The third (novel) view is constructed in a similar manner with a rotation around the unit vector $(0.7, 0.7, 0.14)$ by an angle of 17 degrees. Figure 3.7 (first row) displays the three views. Also in Figure 3.7 (second row) we show the result of applying the transformation due to the four coplanar points $p_1, ..., p_4$ (Step 1, see Section 3.5.1) to all points in the first view. We see that all the coplanar points are aligned with their corresponding points in the second view, and all other points are situated along epipolar lines. The display on the right in the second row shows the final re-projection result (8-point and 6-point methods produce the same result). All points

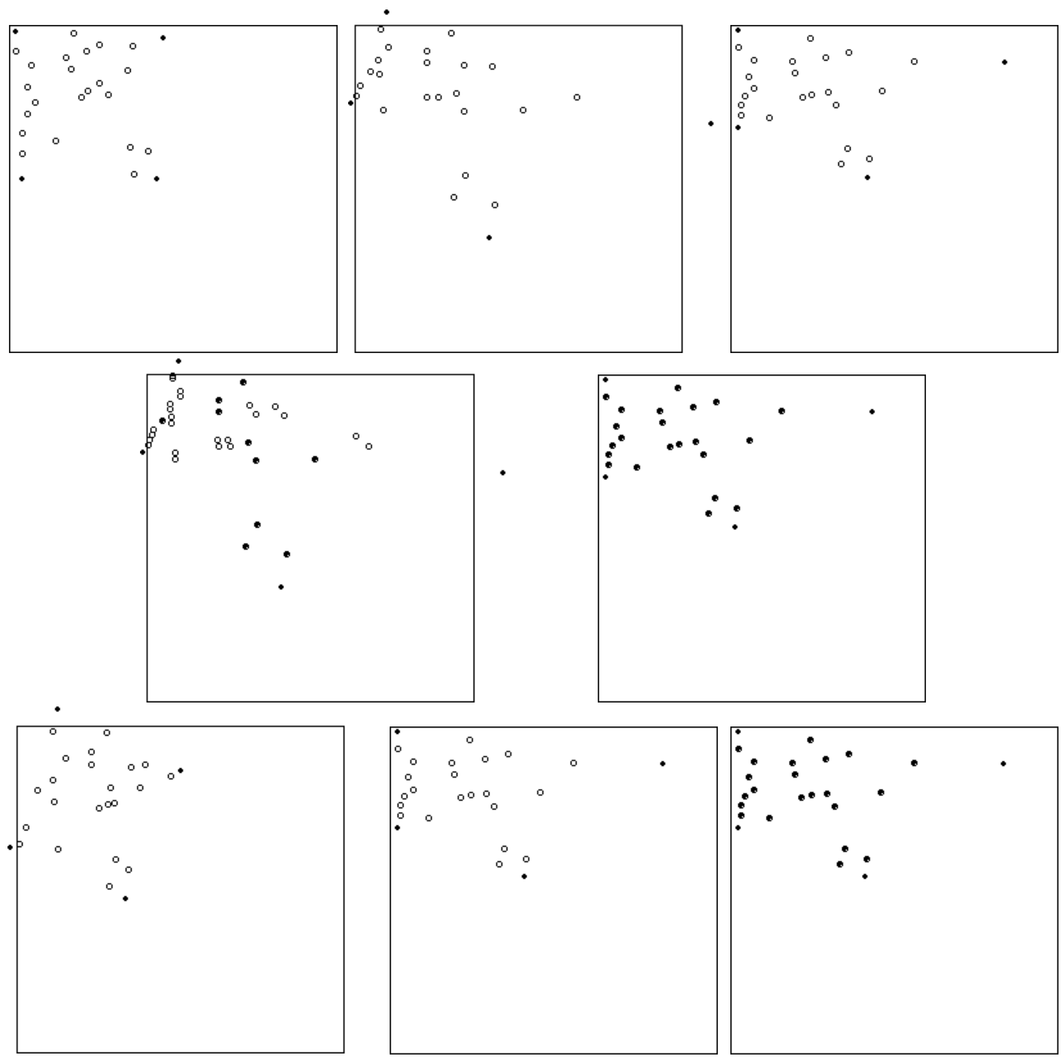
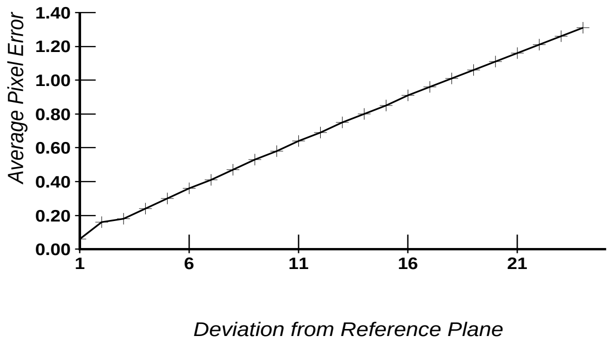


Figure 3.7: Illustration of Re-projection. *Row 1 (left to right):* Three views of the object, two model views and a novel view, constructed by rigid motion following perspective projection. The filled dots represent $p_1, ..., p_4$ (coplanar points). *Row 2:* Overlay of the second view and the first view following the transformation due to the reference plane. All coplanar points are aligned with their corresponding points, the remaining points are situated along epipolar lines. The righthand display is the result of re-projection — the re-projected image perfectly matches the novel image (noise-free situation). *Row 3:* The lefthand display shows the second view which is now orthographic. The middle display shows the third view which is now a perspective projection onto a tilted image plane. The righthand display is the result of re-projection which perfectly matches the novel view.



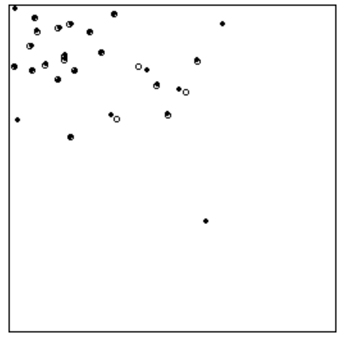


Figure 3.8: Deviation from coplanarity: average pixel error due to translation of P_1 along the optical axis from $z = 100$ to $z = 125$, by increments of one unit. The result of re-projection (overlay of re-projected image and novel image) for the case $z = 125$. The average error is 1.31 and the maximal error is 7.1.

re-projected from the two model views are accurately (noise-free experiment) aligned with their corresponding points in the novel view.

The third row of Figure 3.7 illustrates a more challenging imaging situation (still noise-free). The second view is orthographically projected (and scaled by 0.5) following the same rotation and translation as before, and the novel view is a result of a central projection onto a tilted image plane (rotated by 12 degrees around a coplanar axis parallel to the x-axis). We have therefore the situation of recognizing a non-rigid perspective projection from a novel viewing position, given a rigid perspective projection and a rigid orthographic projection from two model viewing positions. The 6-point re-projection scheme was applied with the result that all re-projected points are in accurate alignment with their corresponding points in the novel view. Identical results were observed with the 8-point algorithms.

The remaining experiments, discussed in the following sections, were done under various noise conditions. We conducted three types of experiments. The first experiment tested the stability under the situation where $P_1, ..., P_4$ are non-coplanar object points. The second experiment tested stability under random noise added to all image points in all views, and the third experiment tested stability under the situation that less noise is added to the privileged six points, than to other points.

3.10.1 Testing Deviation from Coplanarity

In this experiment we investigated the effect of translating P_1 along the optical axis (of the first camera position) from its initial position on the reference plane ($z = 100$) to the

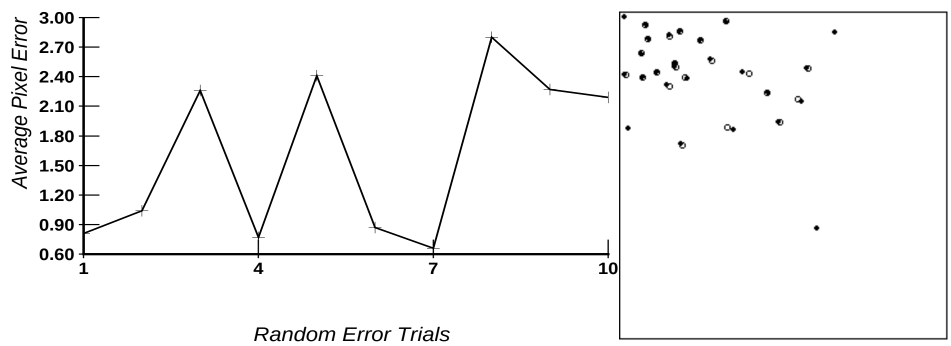


Figure 3.9: Random noise added to all image points, over all views, for 10 trials. Average pixel error fluctuates around 1.6 pixels. The result of re-projection on a typical trial with average error of 1.05 pixels, and maximal error of 5.41 pixels.

farthest depth position ($z = 125$), in increments of one unit at a time. The experiment was conducted using several objects of the type described above (the six privileged points were fixed, the remaining points were assigned random positions in space in different trials), undergoing the same motion described above (as in Figure 3.7, first row). The effect of depth translation to the level $z = 125$ on the location of p_1 is a shift of 0.93 pixels, on p'_1 is 1.58 pixels, and on the location of p''_1 is 3.26 pixels. Depth translation is therefore equivalent to perturbing the location of the projections of P_1 by various degrees (depending on the 3D motion parameters).

Figure 3.8 shows the average pixel error in re-projection over the entire range of depth translation. The average pixel error was measured as the average of deviations from the re-projected point to the actual location of the corresponding point in the novel view, taken over all points. Figure 3.8 also displays the result of re-projection for the case where P_1 is at $z = 125$. The average error is 1.31, and the maximal error (the point with the most deviation) is 7.1 pixels. The alignment between the re-projected image and the novel image is, for the most part, fairly accurate.

3.10.2 Situation of Random Noise to all Image Locations

We next add random noise to all image points in all three views (P_1 is set back to the reference plane). This experiment was done repeatedly over various degrees of noise and over several objects. The results shown here have noise between 0–1 pixels randomly added to the x and y coordinates separately. The maximal perturbation is therefore $\sqrt{2}$, and because the direction of perturbation is random, the maximal error in relative location

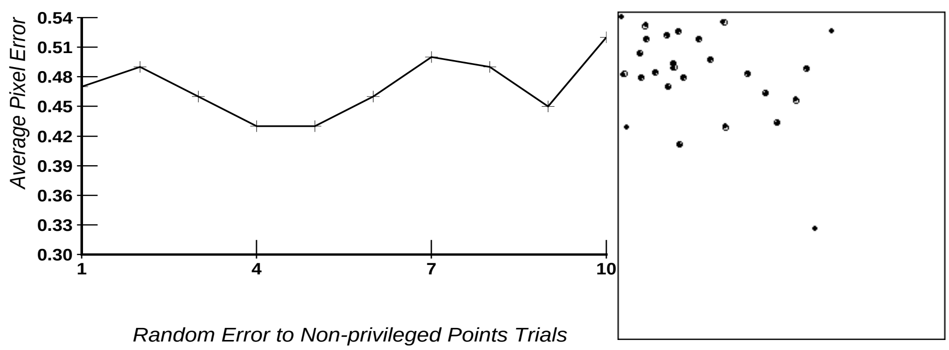


Figure 3.10: Random noise added to non-privileged image points, over all views, for 10 trials. Average pixel error fluctuates around 0.5 pixels. The result of re-projection on a typical trial with average error of 0.52 pixels, and maximal error of 1.61 pixels.

is double, i.e., 2.8 pixels. Figure 3.9 shows the average pixel errors over 10 trials (one particular object, the same motion as before). The average error fluctuates around 1.6 pixels. Also shown is the result of re-projection on a typical trial with average error of 1.05 pixels, and maximal error of 5.41 pixels. The match between the re-projected image and the novel image is relatively good considering the amount of noise added.

3.10.3 Random Noise Case 2

A more realistic situation occurs when the magnitude of noise associated with the privileged six points is much lower than the noise associated with other points, for the reason that we are interested in tracking points of interest that are often associated with distinct intensity structure (such as the tip of the eye in a picture of a face). Correlation methods, for instance, are known to perform much better on such locations, than on areas having smooth intensity change, or areas where the change in intensity is one-dimensional. We therefore applied a level of 0–0.3 perturbation to the x and y coordinates of the six points, and a level of 0–1 to all other points (as before). The results are shown in Figure 3.10. The average pixel error over 10 trials fluctuates around 0.5 pixels, and the re-projection shown for a typical trial (average error 0.52, maximal error 1.61) is in relatively good correspondence with the novel view. With larger perturbations at a range of 0–2, the algorithm behaves proportionally well, i.e., the average error over 10 trials is 1.37.

3.11 Summary of Part I

Although our main interest is achieving recognition via alignment methods, we have chosen to approach the recognition problem from a structure from motion point of view. In Chapter 2 we have shown that the relatively simple domain of parallel projection contains the basic concepts which enable the extension to central projection. First, there is the issue of representation of structure. Second, is the issue of what geometric information is necessary for recognition, and thirdly, is the connection between the geometric approach and the more direct approach obtained by the linear combination of views result.

In this chapter, we first motivated our approach by specifying the major problems in classic approaches for recovering metric structure from two perspective views. We mentioned three problems: (i) critical dependence on an orthographic or perspective model of projection, (ii) internal camera calibration, and (iii) the problem of stereo-triangulation. A necessary approach for dealing with these problems is to work with a non-metric model, but that alone is not sufficient. We argued that the decision of what structure representation to use is of major importance, and, for instance, the representation of structure by coordinates would provide only a partial solution, namely, only a solution to the internal calibration problem.

We have introduced a geometric invariant which we call projective structure, that leads to a system for recovering a relative non-metric shape measurement that does not require internal camera calibration, does not involve full reconstruction of shape (Euclidean or projective coordinates), and treats parallel and central projection as an integral part of one unified system. We have also shown that the invariant can be used for the purposes of visual recognition within the framework of the alignment approach to recognition.

We have shown that the difference between the affine and projective case lie entirely in the location of epipoles, i.e., given the location of epipoles both the affine and projective structure are constructed from the same information captured by four corresponding points projected from four non-coplanar points in space. Therefore, the additional corresponding points in the projective case are used solely for recovering the location of epipoles.

We have shown that the location of epipoles can be recovered under both parallel and central projection using six corresponding points, with the assumption that four of those points are projected from four coplanar points in space, or alternatively by having eight corresponding points without assumptions on coplanarity. The overall method for reconstructing projective structure and achieving re-projection was referred to as the 6-point and the 8-point algorithms. These algorithms have the unique property that projective structure can be recovered from both orthographic and perspective images from uncalibrated cameras. This property implies, for instance, that we can perform recognition of a perspective image of an object given two orthographic images as a model. It also implies greater stability because the size of the field of view is no longer an issue in the process of reconstructing shape or performing re-projection.

Part II

Photometry: Visual Recognition Under Changing Illumination

Previous Approaches and the Problem of Representation

Chapter 4

In this part of the thesis we pursue another aspect of the relation between 3D objects and their images — the photometric aspect, i.e., the relation between objects and their images under changing illumination conditions. Like the geometric mapping from 3D to 2D, we view the photometric aspect as a source of variability that is directly relevant to visual recognition. Most of the research in visual recognition has focused on the geometric aspect of the problem, while leaving the photometric aspect in the hands of data-driven processes that proceed independently of subsequent recognition levels. Photometric issues have, therefore, mostly been dealt with in the context of lower-level processes such as edge detection, lightness and color constancy, and shape from shading.

In this chapter we are concerned with two questions. First and foremost, is it necessary to have a model-based approach to the photometric aspect of recognition? Consider the geometric problem of compensating for changing viewing positions. The major critique of the early approaches, such as those that look for features that are invariant to changing viewing positions or those that attempt to recover generic geometric parts from the input image, is that they are mostly appropriate for dealing with relatively simple objects, like polyhedrons or machine parts (Section 1.3). Alignment model-based methods, on the other hand, were motivated in part by the notion that with complex objects like a face, a horse, a shoe, a loaf of bread, and so forth, a direct coupling between the image and the model at the time of recognition may be more appropriate for dealing with the geometric problem of changing viewing positions. Our first question, therefore, is whether a similar situation applies to the photometric domain, i.e., whether current approaches are mostly appropriate only to relatively simple objects. We address this question by closely examining the current available approaches and by making empirical observations related to human vision (see also Section 1.4.3).

The second question we are concerned with is that of image representation. Assuming that we are pursuing a model-based alignment approach to the photometric problem, then what is the level of image information that needs to be extracted in order to cancel the effects of changing illumination? Consider again the analogy between the photometric and geometric problems. The alignment approach requires us to identify corresponding features between the input image and model images. In the absence of photometric changes between the different views of the object, any set of clearly distinguishable points can be used as features (no labeling is necessary), and in particular the step edges in the light intensity distribution can be used for purposes of representation. The question, therefore, is whether a reduced image representation, such as edges, is sufficient for compensating for illumination changes during recognition, or whether a higher degree of information is necessary, such as the full light intensity distribution. We address this question by examining empirical evidence available from human vision that suggests that in some cases edges alone are not sufficient for visual interpretation, but slightly more than that is sufficient.

In the following sections we first examine the previous approaches for dealing with the problem of changing illumination to see whether they provide a general solution to the problem, or whether a model-based alignment approach is required. We then pursue the question of image representation within the context of a model-based alignment approach.

4.1 Current Approaches

The approaches we review below include computational components of visual analysis that contain, directly or indirectly, a role for illumination. These include edge detection, lightness and color constancy, shape from shading (SFS), and photometric stereo. We will narrow the discussion by assuming that surfaces of interest are matte (Lambertian) or approximately matte.

4.1.1 Edge Detection

The most dominant approach to the problem of changing illumination is to recover features from the image that are invariant to changes of illumination. The key idea is that abrupt changes in intensity provide a sufficiently rich source of features that capture the important aspects for subsequent image analysis, yet at a considerably reduced size. The best known example of such features are step edges, i.e., contours where the light intensity changes relatively abruptly from one level to another. Such edges are often associated with object boundaries, changes in surface orientation, or material properties (Marr 1976, Marr & Hildreth 1980). Edge images contain most of the relevant information in the original grey-level image in cases where the information is mostly contained in changing surface

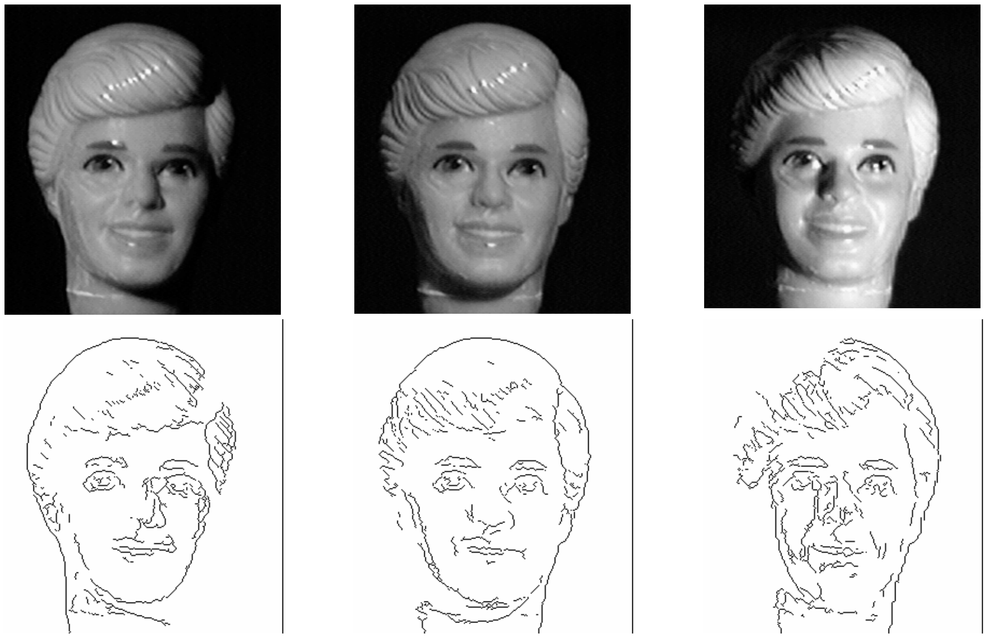


Figure 4.1: Grey-scale images of 'Ken' taken from three different illumination conditions. The bottom row shows the step edges detected by local energy measures followed by hysteresis (Freeman 1992). The step edges look very similar to the ones produced by Canny's edge detection scheme.

material, in sharp changes in surface depth and/or orientation, and in surface texture, color, or greyness. In terms of 3D shape, these are characteristics of relatively simple objects. Therefore, the edges of simple objects are relatively informative (or recognizable) and will change only slightly when the illumination conditions change.

Many natural objects have a more complex structure, however: surface patches do not change orientation abruptly but rather smoothly. In this case, step edges may not be an ideal representation for two reasons: the edge image may not necessarily contain most of the relevant information in the grey-level image, and not all edges are stable with respect to changing illumination. For example, edges that correspond to surface inflections in depth are actually "phantom" edges and depend on the direction of light source (Moses, 1989).

Alternative edge detectors prompted by the need for more recognizable or more stable contour images search instead for extremal points of the light intensity distribution, known as valleys and ridges, or build up a "composite" edge representation made out of the union of step edges, valleys, and ridges (Pearson, Hanna & Martinez 1986, Morrone & Burr 1988, Moses 1989, Perona & Malik 1990, Freeman & Adelson 1991). The composite edge images do not necessarily contain the subset of edges that are stable against changing illumination; they generally look better than step edges alone, but that varies considerably depending on the specific object.

The process of edge detection, producing step edges, ridges, valleys, and composite edge images, is illustrated in Figures 4.1 and 4.2. In Figure 4.1 three 'Ken' images are shown, each taken under a distinct illumination condition, with their corresponding step edges. In Figure 4.2 the ridges, valleys, and the composite edge images of the three original images are shown (produced by Freeman and Adelson's edge and line detector). These results show the invariance of edges are not complete; some edges appear or disappear, some change location, and spurious edges result from shadows (especially attached shadows), specularities, and so forth.

The 'Ken' images and their edge representations also demonstrate the practical side of the problem of recognition under changing illumination conditions. The images appear different to the degree that a template match between any two of them is not likely to succeed without first compensating for the changing illumination.

4.1.2 Recovering Intrinsic Surface Properties: Lightness Constancy

Another possible approach is to decouple the illumination from the image formation equations and thereby recover the surface reflectance, also known as albedo, from the image of the surface. Since surface reflectance is an intrinsic property of the object, we can therefore achieve invariance under changing illumination.

The fact that something like this is possible comes from empirical evidence on human vision. As mentioned in Section 1.4.3, most of the empirical evidence related to the ability of humans to factor out the illumination in judgments of surface color or greyness from a single image, are based on either extremely simple objects, such as planes, or other simple objects such as polyhedrons and cylinders (Land & McCann 1971, Gilchrist 1979, Knill and Kersten 1991). Computational approaches to the problem of recovering surface albedo from 3D objects are also limited to relatively simple objects, such as polyhedrons (Sinha, 1992).

4.1.3 Shape from Shading

Another possible approach is to use the illumination (either assumed or recovered) in order to recover surface structure from the image of the surface. This is known in computer

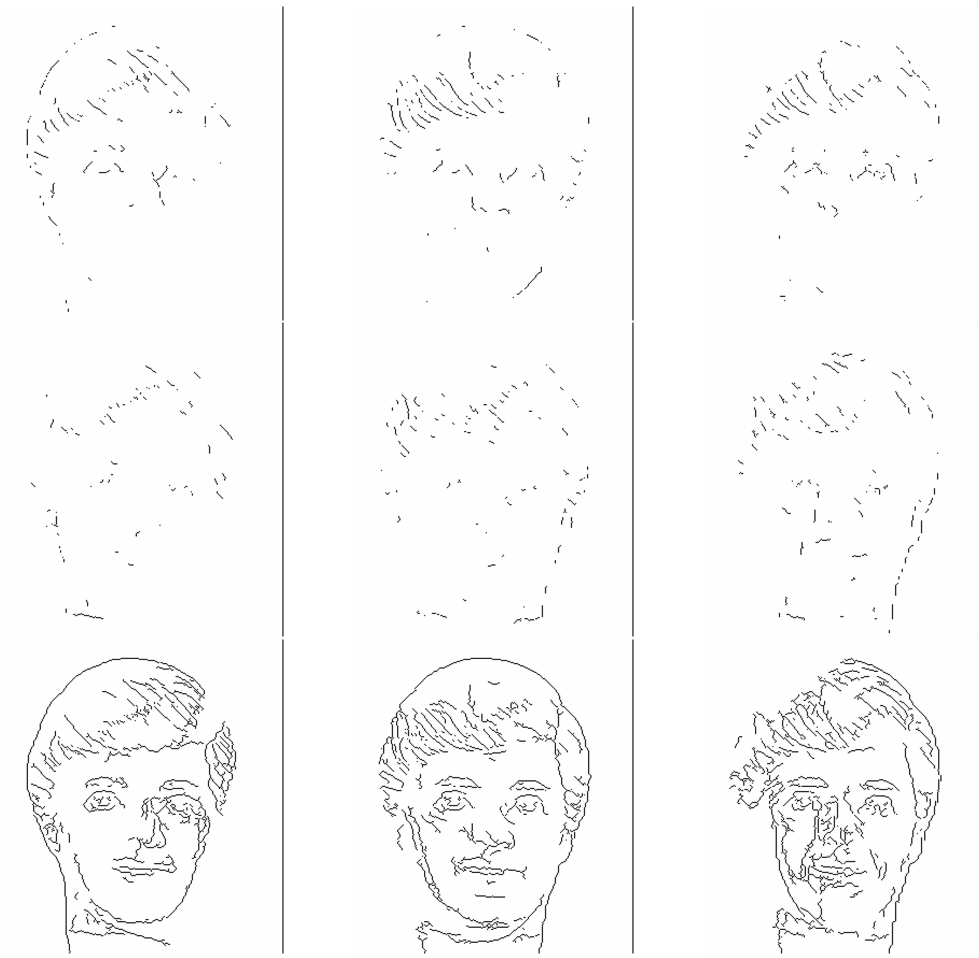


Figure 4.2: Valleys, ridges, and composite contour images produced by Freeman's contour detection method applied to the three images of the previous figure.

vision as "shape from shading". Unlike the previous two approaches, shape from shading methods are often applied to general complex objects, rather than simple objects. However, as described below, there are other restrictions and assumptions that make this approach an unlikely primary vehicle for purposes of recognition.

One class of methods, pioneered by Horn and collaborators (Horn 1977, Ikeuchi & Horn 1981, Horn & Brooks 1986), uses integration techniques for using image grey-values to solve for shape. These methods proceed by propagating constraints from boundary conditions (such as from smooth occluding contours). The drawback of these methods is that they require considerable a priori information, and assumptions, about the scene. These often include surface orientation along surface boundary, and the assumption of uniform albedo (surface is of uniform greyness throughout the area of analysis). In addition, the direction of light source must be recovered. Early techniques (Horn, 1977) required knowledge of light source direction, but more recent schemes (Brooks & Horn, 1985) solve for direction of light source by iterating and interleaving the estimate of shape and light source. Brooks and Horn show that their scheme works well for synthetic images of simple objects like a hemisphere or a cylinder; however, the question of robustness for more general shapes and for real images has remained open.

The second class of methods, pioneered by Pentland (1984), relies on local analysis of shape. These methods do not require knowledge of boundary information, but they do assume uniform surface albedo. In addition, local methods assume that the surface is locally umbilical, but this assumption strictly holds only for a sphere. Nevertheless, local methods may produce good approximations for approximately spherical surfaces. The problem of recovering the direction of light source requires an additional assumption that surface orientation is uniformly distributed over the object (Pentland, 1982).

To conclude, the assumptions and requirements, especially the assumption of uniform albedo, make shape from shading methods an unlikely primary vehicle for purposes of recognition. These limitations do not rule out the possibility of using SFS methods as a component in a recognition system, but this remains an open question.

4.1.4 Photometric Stereo

Another possible approach is to use multiple images of the surface, taken under different illumination conditions, in order to recover intrinsic surface properties such as structure and albedo. This method is known as "photometric stereo" and was pioneered by Woodham (1980). Although photometric stereo belongs to the family of shape from shading methods, we distinguish it here as a separate approach mainly because it is the only approach that can be considered as model-based from the standpoint of achieving recognition.

The basic method is extremely simple, and goes as follows: assume we have three images, I_1, I_2, I_3, of the object taken under three different light source directions. The image intensity at location p in the three images has the form

$$I_j(p) = \mu_p(\boldsymbol{n}_p \cdot \boldsymbol{s}_j) \qquad j = 1, 2, 3,$$

where μ_p is a scalar that represents a mixture of the surface albedo, the spectral response of the image filters, and the spectral composition of light sources — all which is assumed to be fixed for the three images (see Appendix D, for details on image formation and the image irradiance equation). The unit vector $\boldsymbol{s}_j$ is in the direction of the light source, and $\boldsymbol{n}_p$ is the unit vector in the direction of the surface normal. Assuming that we know the directions $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3$, then the three image irradiance equations provide a linear system of three equations and three unknowns per image location ($\mu_p \boldsymbol{n}_p$ is an unknown vector in 3D). The system has a unique solution, provided that the light source directions are linearly independent (no two of them are in the same direction, and the three of them are not coplanar). Photometric stereo, therefore, does not require boundary information, makes no assumptions about surface shape and surface albedo, but requires knowledge of directions and intensity of light source (or only direction if we assume that intensity of light source remains fixed). Knowing the direction of light source puts a heavy burden on the applicability of this method outside the domain of dark-room environments; also, in practice more than three images would be required in order to deal with sensor noise and deviations from the Lambertian model (a fourth image would provide four equations with three unknowns, thereby making it possible to find a least squares solution).

More recent progress in photometric stereo shows that the directions of light sources can be recovered up to an arbitrary rotation of the coordinate system in space. The requirement of knowing the direction of light sources can, as a result, be traded off with the uniform albedo assumption (Woodham, Iwahori and Barman 1991). The method proceeds as follows: let I_p be a vector $(I_1(p), I_2(p), I_3(p))^t$ containing the image intensities at location p of the three images. Because albedo is assumed to be uniform, we can attribute it to the light source vectors $\boldsymbol{s}_j$ (representing light source intensity). We therefore have that $I_p = S\boldsymbol{n}_p$, where S is a 3×3 matrix whose rows are $\boldsymbol{s}_j$. Since $\boldsymbol{n}_p$ is a unit vector, we have

$$I_p^t S^{-t} S^{-1} I_p = I_p^t A I_p = 1,$$

where A is symmetric and positive definite. Therefore, six points are sufficient to determine the six parameters of A (more than six points will define a least squares solution to A). Once A is solved for, we can recover S^{-1}, and hence S, up to a product with an orthogonal matrix. Though the method is elegant and simple, the uniform albedo assumption is too restricting for the improved version of photometric stereo to be of general use for recognition.

In conclusion, we have seen that purely data-driven approaches such as edge detection, lightness constancy, and shape from shading are either limited to relatively simple objects, or make several restricting assumptions regarding surface greyness, illumination, and shape.

Figure 4.3: Mooney faces and their level-crossings.

The only model-based approach, the photometric stereo method, is also limited to relatively simple cases, not necesserily simple objects, of surfaces with uniform greyness.

4.2 The Question of Image Representation

Subsequent to motivating our pursuit after a model-based alignment approach to the photometric problem of recognition, we turn our attention to the question of image representation. The question of image representation is, what kind of image features are necessary for compensating for changing illumination? The question of representation is motivated by two considerations. The first is phenomenological. It appears that in some cases in human vision, more than edges are required for visual interpretation, but not much more. We would like to examine this situation and make a connection with the computational results presented in the next two chapters. The second consideration is a practical one. The less we rely on the exact light intensity distribution in the image, and rely more on reduced representations of the image, the more stable the overall scheme would be. Image grey-values are prone to errors because of poor illumination, low signal to noise ratio due to distance to the viewer, and other effects that may occur at recognition time and cause fluctuations in image intensities. These situations may have a lesser effect if instead more

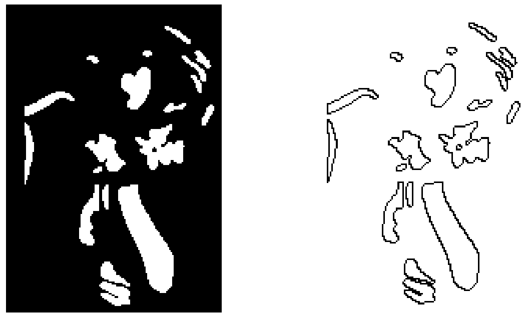


Figure 4.4: A less interpretable Mooney picture and its level-crossings

compact representations are used.

It appears that in some cases in human vision the interpretation process involves more than just contours. A well-known example is the set of thresholded face images produced by Mooney (1960) for clinical recognizability tests, known as the closure faces test, in which patients had to sort the pictures into general classes that include: boy, girl, grown-up man or woman, old man or woman, and so forth. An example of Mooney's pictures are shown in Figure 4.3. Most of the control subjects could easily label most of the pictures correctly. Some of Mooney's pictures are less interpretable (for example, Figure 4.4), but as a general phenomenon it seems remarkable that a vivid visual interpretation is possible from what seems an ambiguous collection of binary patches that do not bear a particularly strong relationship to surface structure or other surface properties.

Mooney images are sometimes referred to as representing the phenomenon of "shape from shadows" (Cavanagh 1990). Although some Mooney images do contain cast shadows, the phenomenon is not limited to the difficulty of separating shadow borders from object contours. The thresholded image shown in Figure 4.5, for example, is not less difficult to account for in computational terms, yet the original image was not lit in a way to create cast or attached shadows.

In Section 1.4.3 we used Mooney-kind images to argue in favor of a model-based computation in which illumination is compensated for during recognition. Here we wish to point out another aspect of these kinds of images. It is evident that the contours (level-crossings) alone are not interpretable, as can be seen with the original Mooney pictures and with the level-crossing image in Figure 4.5. It seems that only when the distinction of what regions are above the threshold and what are below the threshold is made clear (we refer to that as adding "sign-bits") does the resulting image become interpretable. This appears to be true not only for thresholded images but also for step edges and their sign-bits (see Figure 4.5,

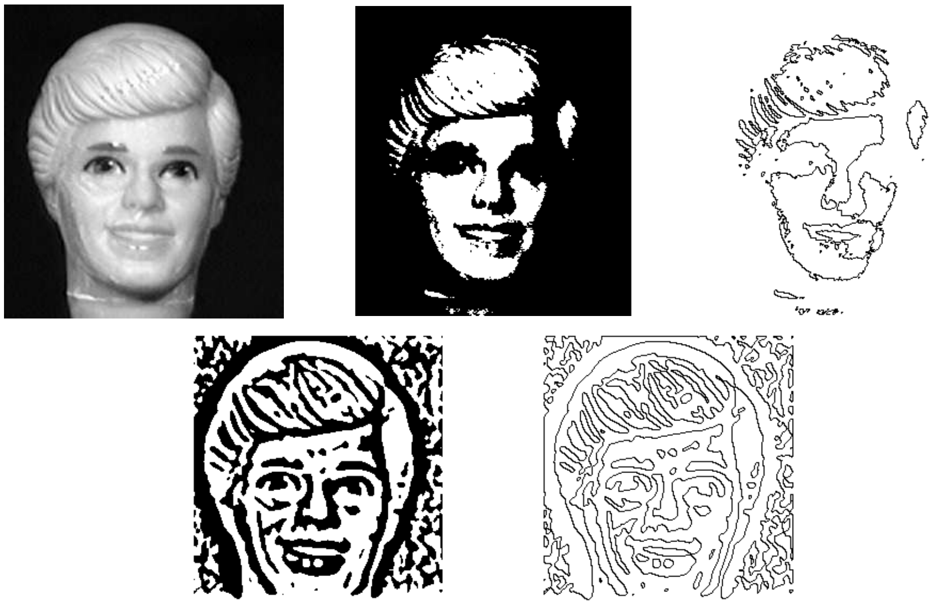


Figure 4.5: *Top Row:* A 'Ken' image represented by grey-levels, the same image followed by a threshold, the level-crossings of the thresholded image. The thresholded image shown in the center display is difficult to account for in computational terms, yet the original image was not lit in a way to create cast or attached shadows. *Bottom Row:* The sign-bits of the Laplacian of Gaussian operator applied to the original image, and its zero-crossings (step edges). Interpretability of the sign-bit image is considerably better than the interpretability of the zero-crossings.

bottom row).

It appears, therefore, that in some cases in human vision the illumination is factored out within the recognition process using top-down information and that the process responsible apparently requires more than just contours — but not much more. We refer from here on to the Mooney-kind of images as reduced images. From a computational standpoint we will be interested not only in factoring out the illumination, in an alignment model-based appraoch, but also in doing so from reduced images — this is described in the next two chapters.

4.3 Summary

The two central questions addressed in this chapter are, is there a need for a model-based approach for dealing with the effects of illumination in recognition? And what are the limits on image information in order to make that process work? We arrive at two conclusions. First, image properties alone do not appear to be sufficient for obtaining a general result of factoring out the illumination prior to recognition. This conclusion is based partly on empirical observations resulting from Mooney-kind of images (also in Section 1.4.3) and partly on observing the limitations of various possible approaches in the areas of edge detection, lightness constancy, and shape from shading. Second, the process responsible for factoring out the illumination during the recognition process appears to require more than just contour information, but just slightly more. We refer to what seems a necessary input level as reduced images. Although reduced images are not a representative of natural input images, they are, nevertheless, an especially difficult type of inputs that humans are able to interpret very well. It may be also possible to view reduced images as an extreme case of a wider phenomenon of interpreting low quality images, such as images seen in newspapers, images produced by photo-copying the original, or images taken under poor illumination conditions. These types of inputs, however, do not fall within the scope of this thesis.

Photometric Alignment

Chapter 5

A model-based alignment type of approach to the photometric problem assumes that we have stored several images of the object, taken under distinct illumination conditions, with which a compensating transformation can be found. This transformation is such that the model (represented by the model images) can be converted to match the novel input image of the object. We have seen this done in several ways in the geometric domain, and also in the photometric domain with the use of photometric stereo. With photometric stereo, however, we had to assume knowledge of light source parameters, or to assume that the surface is of uniform albedo — both of which we want to avoid. Our approach is similar to the linear combination of views (Ullman & Basri, 1989) that was introduced in the geometric domain. In other words, we define a direct algebraic connection between all images of a matte surface under changing illumination conditions. The alignment scheme that makes use of this result for purposes of recognition is referred to as the photometric alignment scheme. The photometric problem was defined earlier in Section 1.2, and is re-produced below:

Photometric Problem: *We are given three images of an arbitrary convex matte surface. The images are taken under three different arbitrary settings of point light sources. For any arbitrary image determine whether the image can be produced by the surface under some illumination condition.*

The photometric problem assumes the surface is convex in order to avoid the problem of cast shadows (this assumption is implicitly contained in the photometric stereo method as well). Although our analytic results strictly hold only under the conditions specified above, these restrictions are not critical in practice. The main results derived in this chapter include the following:

- The effect of changing direction and intensity of light sources can be factored out by a linear combination of images of the same object.

- The photometric alignment scheme can be used to detect specular regions in the image.
- The effects of changing the spectral composition of light sources can be factored out by linearly combining the color bands of a single model image.

5.1 The Linear Combination of Grey-scale Images

Definition 4 *An order k Linear Reflectance Model is defined as the scalar product $\boldsymbol{x} \cdot \boldsymbol{a}$, where $\boldsymbol{x}$ is a vector in k-dimensional Euclidean space of invariant surface properties (such as surface normal, surface albedo, and so forth), and $\boldsymbol{a}$ is an arbitrary vector (of the same dimension).*

The Lambertian model of reflection is an obvious case of an order 3 linear reflectance model. As described in more detail in Appendix D, the grey-value, $I(p)$, at location p in the image can be represented by the scalar product of the surface normal vector and the light source vector,

$$I(p) = \boldsymbol{n}_p \cdot \boldsymbol{s}.$$

Here the length of the surface normal $\boldsymbol{n}_p$ represents the surface albedo (a scalar ranging from zero to one). The length of the light source vector $\boldsymbol{s}$ represents a mixture of the spectral response of the image filters, and the spectral composition of light sources — both of which are assumed to be fixed for all images of the surface (we assume for now that light sources can change direction and level of intensity but not spectral composition).

Another example of a linear reflectance model is the image irradiance of a tilted Lambertian surface under a hemispherical sky. Horn (1986, pp. 234) shows that the image irradiance equation is $E\delta_p cos^2 \frac{\alpha}{2}$, where α is the angle between the surface normal and the zenith, E is the intensity of light source, and δ_p is the surface albedo. The equation is an order 4 linear reflectance function:

$$I(p) = \frac{1}{2} E\delta_p (1 + cos\alpha) = \boldsymbol{n}_p \cdot \boldsymbol{s} + |\boldsymbol{n}_p| \cdot |\boldsymbol{s}| = (\boldsymbol{n}_p, |\boldsymbol{n}_p|)^t (\boldsymbol{s}, |\boldsymbol{s}|),$$

where $\boldsymbol{s}$ represents the direction of zenith, whose length is $\frac{E}{2}$.

Proposition 6 *An image of an object under an order k linear reflection model $I(p) = \boldsymbol{x}(p) \cdot \boldsymbol{a}$ can be represented as a linear combination of a fixed set of k images of the object.*

Proof: Let $\boldsymbol{a}_1, ..., \boldsymbol{a}_k$ be some arbitrary set of basis vectors that span k-dimensional Euclidean space. The image intensity $I(p) = \boldsymbol{x}(p) \cdot \boldsymbol{a}$ is therefore represented by

$$I(p) = \boldsymbol{x}(p)[\alpha_1 \boldsymbol{a}_1 + ... + \alpha_k \boldsymbol{a}_k] = \alpha_1 I_1(p) + ... + \alpha_k I_k(p),$$

where $\alpha_1, ..., \alpha_k$ are the linear coefficients that represent $\boldsymbol{a}$ with respect to the basis vectors, and $I_1, ..., I_k$ are the k images $I_k(p) = \boldsymbol{x}(p) \cdot \boldsymbol{a}_k$. $\square$

To see the relevance of this proposition to visual recognition, consider the case of a Lambertian surface under a point light source (or multiple point light sources). Assume we take three pictures of the object I_1, I_2, I_3 from light source directions $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3$, respectively. The linear combination result is that any other image I of the object, taken from a novel setting of light sources, is simply a linear combination of the three pictures,

$$I(p) = \alpha_1 I_1(p) + \alpha_2 I_2(p) + \alpha_3 I_3(p),$$

for some coefficients $\alpha_1, \alpha_2, \alpha_3$ (this observation was made independently by Yael Moses). The coefficients can be solved by observing the grey-values of three points providing three equations. Using more than three points will provide a least squares solution. The solution is unique provided that $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3$ are linearly independent, and that the normal directions of the three sampled points span all other surface normals (for a general 3D surface, for example, the three normals should be linearly independent).

Alignment-based recognition under changing illumination can proceed in the following way. The images $I_1, ..., I_k$ are the model images of the object (three for Lambertian under point light sources). For any new input image I, rather than matching it directly to previously seen images (the model images), we first select a number of points (at least k) to solve for the coefficients, and then synthesize an image $I' = \alpha_1 I_1 + ... + \alpha_k I_k$. If the image I is of the same object, and the only change is in illumination, then I and I' should perfectly match (the matching is not necessarily done at the image intensity level, one can match the edges of I against the edges of I', for example). This procedure has factored out the effects of changing illumination from the recognition process without recovering scene information, i.e. surface albedo or surface normal, and without assuming knowledge of direction of light sources (as photometric stereo does). Another property of this method is that one can easily find a least squares solution for the reconstruction of the synthesized image, thereby being less sensitive to errors in the model, or input errors.

In addition to the properties listed above, the photometric alignment approach also shares the general properties of the geometric alignment methods including that: (i) the procedure can be applied whether the image and the model are of the same object or not, (ii) the complexity of the object is not of critical importance, although here it may have an effect by introducing more cast shadows (see below), and (iii) the actual matching is performed in a pictorial manner without the need to recover scene information, and without the application of top-down reasoning processes. We consider next two situations that occur with general surfaces (rather than convex matte surfaces). The first situation described in the next section is that of cast and attached shadows; and the second situation, described in Section 5.1.2, is that of specular reflections arising from non-matte surfaces.

5.1.1 Attached and Cast Shadows

We have assumed that surfaces are convex because the linear combination result requires that points be visible to the light sources. In a general non-convex surface object points may be occluded from some, or from all, the light sources. This situation generally leads to two types of shadows known as attached and cast shadows. A point P is in an attached shadow if the angle between the surface normal and the direction of light source is obtuse ($\boldsymbol{n}_p \cdot \boldsymbol{s} < 0$). An object point P is in a cast shadow if it is obstructed from the light source by another object or by part of the same object. An attached shadow, therefore, lies directly on the object, whereas cast shadows are thrown from one object onto another, or from one part onto another of the same object (such as when the nose casts a shadow on the cheek under oblique illumination).

In the case of attached-shadows, a correct reconstruction of the image grey-value at p does not require that the object point P be visible to the light source $\boldsymbol{s}$, but only that it be visible to the light sources $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3$. If P is not visible to $\boldsymbol{s}$, then the linear combination will produce a negative grey-value (because $\boldsymbol{n}_p \cdot \boldsymbol{s} < 0$), which can be set to 0 for purposes of display or recognition.

If P is not visible to one of the model light sources, say $\boldsymbol{s}_1$, then the linear combination of the three model images is to reconstructing $I'(p)$ under a light source $\boldsymbol{s}'$ which is the projection of $\boldsymbol{s}$ onto the sub-space spanned by $\boldsymbol{s}_2, \boldsymbol{s}_3$. This implies that photometric alignment would perform best in the case where the novel direction of light source $\boldsymbol{s}$ is within the cone of directions $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3$.

The remaining case is when the object point P is in a cast shadow with respect to the novel light direction $\boldsymbol{s}$. In this case there is no way to predict a low, or zero, grey-value for $I'(p)$ and the reconstruction will not match $I(p)$. Therefore, cast shadow regions in the

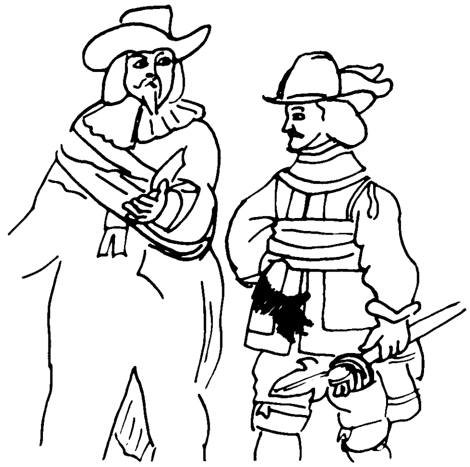


Figure 5.1: Rembrandt's *Night Watch* illustrating that cast shadows may be intellectually understood, yet visually non-obvious. The hand of the figure on the left (the captain) is casting a shadow on the figure on the right (the lieutenant). The shadow is understood as created from the captain's gesticulating hand, but does not appear to have a perceptual connection to the object on which it appears (Arnheim, 1954).

novel image are not modeled in this framework, and hence, the performance degrades with increasing number and extent of cast-shadows in the novel image.

With regard to human vision, there appears to be a marked increase in difficulty in interpreting cast shadows compared to attached shadows. Arnheim (1954) discusses the effect of cast shadows on visual perception, its relation to chiaroscuro in Renaissance art, and its symbolism in various cultures. He points out that cast shadows often interfere with the object's integrity, whereas attached shadows are often perceived as an integral part of the object. Rembrandt's *Night Watch*, displayed in Figure 5.1, is an example of a shadow that is intellectually understood, yet is not visually obvious. Although the shadow is cast upon a different object, the general observation is that the more the cast-shadow extends from the part that throws it, the less meaningful is the connection made with the object. The interpretability of cast shadows is also illustrated by 'Ken' images displayed in Figure 5.2. The three model images have extensive attached shadows that appear naturally integrated with the object. The cast shadow region thrown from the nose in the image on the right appears less integrated with the overall composition of the image.

In conclusion, attached shadows in the novel image, or shadows in general in the model images, do not have significant adverse effects on the photometric alignment scheme. Cast

Figure 5.2: Three model images of a plastic doll, taken under a single light source from three different directions (non-coplanar) and intensities. Note that the surface is not perfectly matte, and the images contain shadows and specular reflections.

shadows in the novel image, cannot be reconstructed or even approximated, and therefore are not modeled in this framework. It may be noted that apparently there is a perceptual difference between attached and cast shadows, whereby the latter may appear to be disconnected from the object upon which they are cast.

5.1.2 Detecting and Removing Specular Reflections

The linear combination result and the photometric alignment scheme that followed assume that objects are matte. In general, inhomogeneous surfaces are dominantly Lambertian, except for isolated regions that are specularly reflecting light (see Appendix D). In practice, if the specular component is ignored, the reconstructed image has the specular regions of all three model images combined together, and the specular regions of the novel image are not reconstructed. For purposes of recognition, as long as the specular regions are relatively small, they do not seem to have a significant adverse effect on the overall photometric alignment scheme. Nevertheless, the alignment method can be used to detect the specular regions and replace them with the Lambertian reflectance provided that four images are used.

The Detection of specular points is based on the observation that if a point is in the specular lobe, then it is likely to be so only in one of the images at most. This is because the specular lobe occupies a region that falls off exponentially from the specular direction. In general we cannot detect the specular points by simply comparing grey-values in one image with the grey-values of the same points in the other images because the intensity of the light source may arbitrarily change from one image to another.

By using Proposition 6, that is, the result that three images uniquely determine the Lambertian component of the fourth image, we can, thereby, compare the reconstructed intensity of the fourth image with the observed intensity, and check for significant deviations. For every point p, we select the image with the highest intensity, call it I_s, and reconstruct $I'_s(p)$ from the other three images (we recover the coefficients once, based on points that are not likely to be specular or shadowed, i.e. do not have an especially high or low intensity). If $I_s(p)$ is in the specular lobe, then $I'_s(p) << I_s(p)$. To avoid deviations that are a result of shadowed points, we apply this procedure to points for which none of the images has an especially low grey-value.

In practice we observe that the deviations that occur at specular points are of an order of magnitude higher than deviations anywhere else, which makes it relatively easy to select a threshold for deciding what is specular and what is not. A similar approach for detecting specular points was suggested by Coleman and Jain (1982) based on photometric stereo. The idea is to have four images and to reconstruct the normal at each point from every subset of three images. If the point in question is not significantly specular, then the reconstructed normals should have the same direction and length, otherwise the point is likely to be specular. Their method, however, requires knowledge of direction and intensity of light sources, whereas in our method we do not.

5.1.3 Experimental Results

We used the three 'Ken' images displayed in Figure 5.2 as model images for the photometric alignment scheme. The surface of the doll is non-convex almost matte which gives rise to specular reflections and shadows. The novel image (shown in Figure 5.3) was taken using light source directions that were within the cone of directions used to create the model images. In principle, one can use novel light source directions that are outside the cone of directions, but that will increase the likelihood of creating new cast shadow regions. The reconstruction was based on a least squares solution using eight points. The points were chosen automatically by searching for smooth regions of image intensity. The search was restricted to the area of the face, not including the background. To minimize the chance of selecting shadowed or specular points, a point was considered as an admissible candidate if it was contained in an 8×8 sized smooth area, and its intensity was not at the low or high end of the spectrum. We then selected eight points that were widely separated from each other. The reconstructed image (linear combination of the three model images) is displayed in Figure 5.3 together with its step edges. The novel and reconstructed image

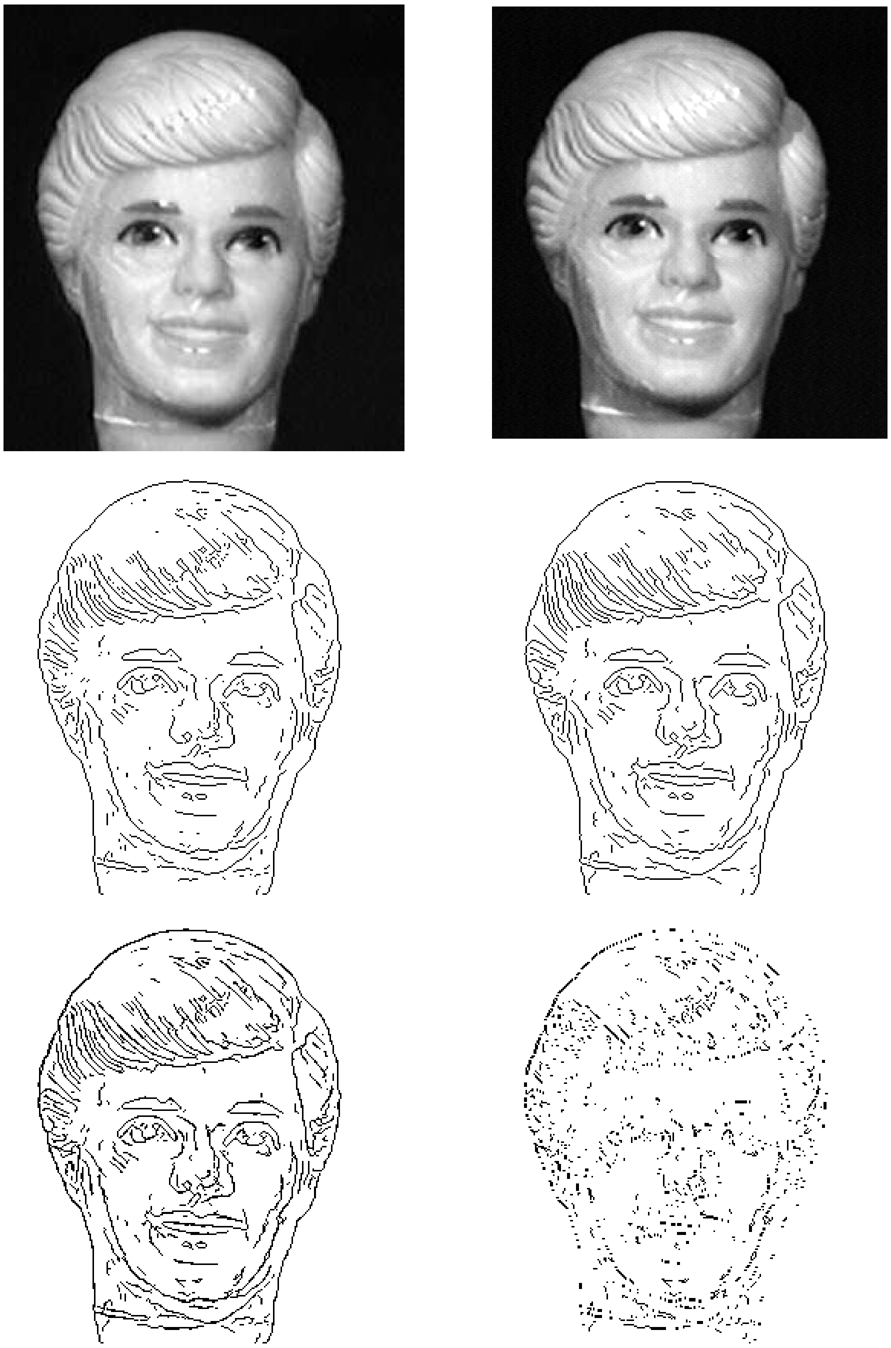


Figure 5.3: Reconstructing a novel image. *Row 1 (left to right):* A novel image taken from two point light sources, and the reconstructed image (linear combination of the three model images). *Row 2:* Step edges of the novel and reconstructed images. *Row 3:* Overlaying both edge maps, and subtracting (*xor* operation) the edge maps from each other. The difference between the images both at the grey-scale and edge level is hardly noticeable.

are visually very similar at the grey-value level, and even more so at the edge-map level. The difference between the two edge maps is negligible and is mostly due to quantization of pixel locations.

In conclusion, this result shows that for the purposes of recognition, the existence of shadows and (small) specular regions in the model images do not have a significantly adverse effect on the reconstruction. Moreover, we did not use a matte surface for the experiment, illustrating the point that plastic surfaces are dominantly Lambertian, and therefore sufficiently applicable to this method.

Figure 5.4 demonstrates the specular detection scheme. The method appears to be successful in identifying small specular regions. Other schemes for detecting specular regions using the dichromatic model of reflection often require a relatively large region of analysis and, therefore, would have difficulties in detecting small specular regions (Shafer 1985, Klinker, Shafer and Kanade 1990).

5.2 The Linear Combination of Color Bands

The photometric problem considered so far involved only changes in direction and intensity of light sources, but not changes in their spectral compositions. Light sources that change their spectral composition are common as, for example, sunlight changes its spectral composition depending on the time of day (because of scattering). The implication for recognition, however, is not entirely clear because there may be an adaptation factor involved rather than an explicit process of eliminating the effects of illumination. Adaptation is not a possibility when it comes to changing direction of light source, because objects are free to move in space and hence change their positions with respect to the light sources. Nevertheless, it is of interest to explore the possibility of compensating for changing spectral composition as well as direction of light sources.

We assume, for reasons that will be detailed below, that our surface is either *neutral*, or is of the same color, but may change in luminosity. A neutral surface is a grey-scale surface only affecting the scale of light falling on the surface, but not its spectral composition. For example, the shades of grey from white to black are all neutral. Note that the assumption is weaker than the uniform albedo assumption because we allow change in luminosity, but is less general than what we had previously because we do not allow changes in hue or saturation to occur across the surface. We also assume that our model of the object

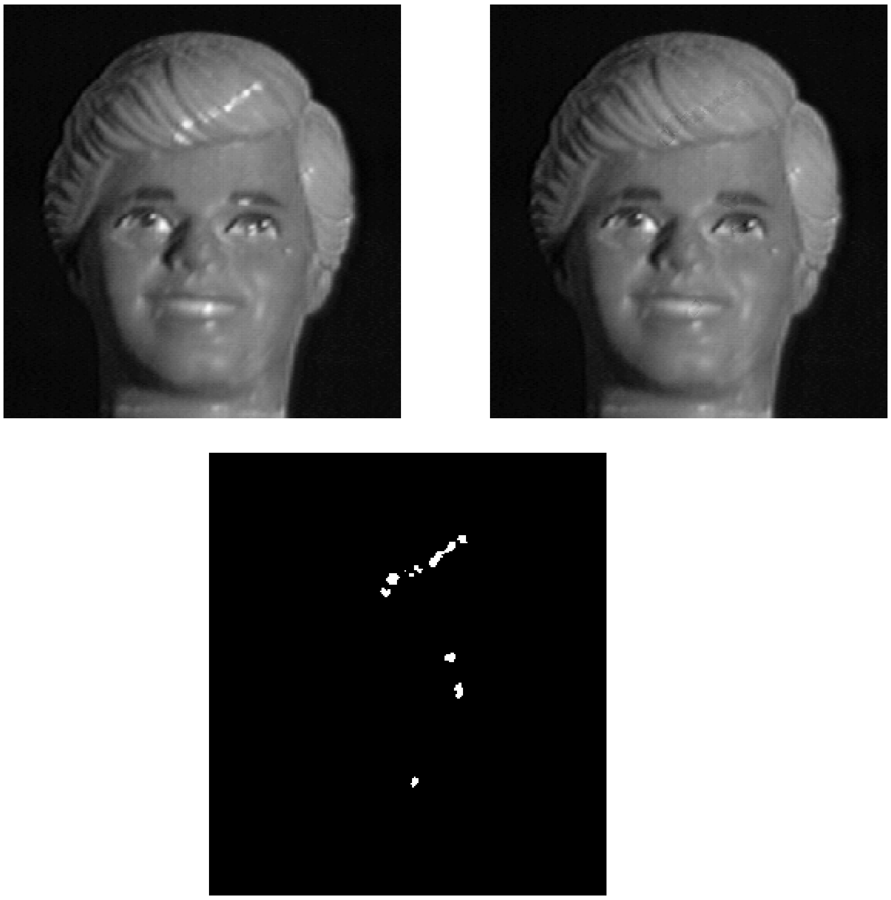


Figure 5.4: Detecting and removing specular regions. *Row 1:* The image on the left is a novel image, and the one on the right is the same image following the procedure for detecting and removing the specular regions. The specular regions are replaced with the reconstructed grey-value from the model images. *Row 2:* The specular regions that were detected from the image.

consists of a single color image obtained by overlaying three color images of the object each taken from a distinct direction of light source having a distinct spectral composition.

Let I_r, I_g, I_b be the three color bands that together define the color picture. Let $\delta_p \rho(\lambda)$ be the surface reflectance function. Note that the neutral surface assumption means that across the surface $\rho(\lambda)$ is fixed, but δ_p may change arbitrarily. Let $S_1(\lambda), S_2(\lambda), S_3(\lambda)$ be the spectral composition of the three light sources, and $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3$ be their directions. As before, we require that the directions be non-coplanar, and that the spectral compositions be different from each other. This, however, does not mean that the three spectral functions should form a basis (such as required in some color constancy models, Maloney and Wandell 1986). Finally, let $R_r(\lambda), R_g(\lambda), R_b(\lambda)$ be the spectral sensitivity functions of the three CCD filters (or cones). The composite color picture (taking the picture separately under each light source, and then combining the results) is, therefore, determined by the following equation:

$$\begin{aligned}\begin{pmatrix} I_r(p) \\ I_g(p) \\ I_b(p) \end{pmatrix} &= \begin{pmatrix} \int S_1(\lambda)\rho(\lambda)R_r(\lambda)d\lambda \\ \int S_1(\lambda)\rho(\lambda)R_g(\lambda)d\lambda \\ \int S_1(\lambda)\rho(\lambda)R_b(\lambda)d\lambda \end{pmatrix} \boldsymbol{n}_p \cdot \boldsymbol{s}_1 + \begin{pmatrix} \int S_2(\lambda)\rho(\lambda)R_r(\lambda)d\lambda \\ \int S_2(\lambda)\rho(\lambda)R_g(\lambda)d\lambda \\ \int S_2(\lambda)\rho(\lambda)R_b(\lambda)d\lambda \end{pmatrix} \boldsymbol{n}_p \cdot \boldsymbol{s}_2 + \\ &\quad \begin{pmatrix} \int S_3(\lambda)\rho(\lambda)R_r(\lambda)d\lambda \\ \int S_3(\lambda)\rho(\lambda)R_g(\lambda)d\lambda \\ \int S_3(\lambda)\rho(\lambda)R_b(\lambda)d\lambda \end{pmatrix} \boldsymbol{n}_p \cdot \boldsymbol{s}_3,\end{aligned}$$

where the length of $\boldsymbol{n}_p$ is δ_p. This can re-written in matrix form, as follows:

$$\begin{aligned}\begin{pmatrix} I_r(p) \\ I_g(p) \\ I_b(p) \end{pmatrix} &= \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \boldsymbol{n}_p \cdot \boldsymbol{s}_1 + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \boldsymbol{n}_p \cdot \boldsymbol{s}_2 + \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} \boldsymbol{n}_p \cdot \boldsymbol{s}_3 \\ &= [\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}] \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \\ \boldsymbol{s}_3 \end{bmatrix} \boldsymbol{n}_p = A\boldsymbol{n}_p.\end{aligned}$$

The 3×3 matrix $[\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}]$ is assumed to be non-singular (for that reason we required that the spectral composition of light sources be different from one another), and therefore the matrix A is also non-singular. Note that because of the assumption that the surface is neutral, the matrix A is independent of position. Consider any novel image of the same surface, taken under a new direction of light source with a possible different spectral composition. Let the novel picture be J_r, J_g, J_b. The red color band, for instance, can be represented as a linear combination of the three color bands I_r, I_g, I_b, as follows:

$$J_r(p) = \left[\int S(\lambda)\rho(\lambda)R_r(\lambda)d\lambda\right]\boldsymbol{n}_p\cdot\boldsymbol{s} = \boldsymbol{n}_p\cdot(\alpha_1 A_1+\alpha_2 A_2+\alpha_3 A_3) = \alpha_1 I_r(p)+\alpha_2 I_g(p)+\alpha_3 I_b(p)$$

where A_1, A_2, A_3 are the rows of the matrix A. Because A is non-singular, the row vectors form a basis that spans the vector $[\int S(\lambda)\rho(\lambda)R_r(\lambda)d\lambda]\boldsymbol{s}$ with some coefficients $\alpha_1, \alpha_2, \alpha_3$. These coefficients are fixed for all points in the red color band because the scale $\int S(\lambda)\rho(\lambda)R_r(\lambda)d\lambda$ is independent of position (the neutral surface albedo δ_p is associated with the length of $\boldsymbol{n}_p$). Similarly the remaining color bands J_g, J_b are also represented as a linear combination of I_r, I_g, I_b, but with different coefficients. We have, therefore, arrived at the following result:

Proposition 7 *An image of a Lambertian object with a neutral surface reflectance (grey-scale surface) taken under an arbitrary point light source condition (intensity, direction and spectral composition of light source) can be represented as a linear combination of the three color bands of a model picture of the same object taken under three point light sources having different (non-coplanar) directions and different spectral composition.*

For a neutral surface, the linear combination of color bands can span only images of the same surface with the same hue and saturation under varying illumination conditions. The combination of color bands of a non-neutral surface spans the space of illumination *and* color (hue and saturation). That is, two surfaces with the same structure but with different hue and saturation levels, are considered the same under the photometric alignment scheme.

5.3 Summary

The photometric alignment scheme presented in this chapter is a model-based approach similar to photometric stereo since multiple images of the same object taken from different illumination conditions are recorded. Unlike photometric stereo, we do not use these images in order to recover intrinsic properties of the object, but rather to directly compensate for the change in illumination conditions for any other novel image of the object. This difference is critical for it enables us to avoid the limitations of photometric stereo by allowing an arbitrary distribution of surface albedo, and by not having to assume or recover the parameters associated with the light sources. We have discussed the situations of shadows and specular reflections. The conclusion was that attached shadows in the model and novel images are generally not a problem for the photometric alignment method, but cast shadows in the novel image are. The alignment scheme, therefore, degrades with increasing cast shadow regions in the novel image. As a result of this, photometric alignment when applied to general non-convex surfaces is most suitable for reconstructing novel images whose illumination conditions are in between those used to create the model images. We have also seen that specular reflections arising from inhomogeneous surfaces can be detected and removed if necessary. Finally, we have extended the basic result to deal with color images and the problem of changing spectral composition of light sources in addition to their directions.

Photometric Alignment with Reduced Images

Chapter 6

The primary purpose of this chapter is to address the question of image representation within the context of the photometric alignment approach. In Chapter 4 we arrived at two conclusions based on empirical observations on human vision: first, it appears that in some cases illumination is factored out during the recognition process in a model-based manner. Second, the process responsible for factoring out the illumination during the recognition process appears to require more than just contour information, but just slightly more. We have addressed the first issue in the previous chapter by proposing the photometric alignment method, which can directly factor out the illumination during the model-to-image matching stage by using the information contained in the grey-values of the model and novel images.

In this chapter we explore the possibilities of using less than grey-values for purposes of factoring out the illumination. In other words, since the photometric alignment method is essentially about recovering the linear coefficients that represent the novel image as a linear combination of the three model images, then the question is whether those coefficients can be recovered by observing more reduced representations of the novel image, such as edges, edges and gradients, sign-bits, and so forth. Specifically, we are most interested in making a computational connection with the empirical observation that sign-bits appear to be sufficient for visual interpretation, whereas edges alone are not. The main results derived in this chapter include the following:

- We show that level-crossing or zero-crossing contours of the novel image are theoretically sufficient for recovering the linear coefficients for the model images. This result requires, however, that contours be given at a sub-pixel accuracy.

- We show that the requirement of accuracy can be traded off by adding the image gradients along the contours.

- The accuracy of edges can be traded off, if instead of edges, the sign-bits are given everywhere. This possibility is shown to be the most appealing computationally and provides a connection to our previous observation that edges alone are in some cases insufficient for visual interpretation, but sign-bits are sufficient.

6.1 Photometric Alignment from Contours

Proposition 8 *The coefficients that span an image I from three model images, as described in proposition 6, can be solved, up to a common scale factor, from just the contours of I, zero-crossings or level-crossings.*

Proof: Let α_j be the coefficients that span I by the basis images I_j, $j = 1, 2, 3$, i.e. $I = \sum_j \alpha_j I_j$. Let f, f_j be the result of applying a Laplacian of Gaussian (LOG) operator, with the same scale, on images I, I_j, $j = 1, 2, 3$. Since LOG is a linear operator we have $f = \sum_j \alpha_j f_j$. Since $f(p) = 0$ along zero-crossing points p of I, then by taking three zero-crossing points, which are not on a cast shadow border and whose corresponding surface normals are non-coplanar, we get a homogeneous set of equations from which α_j can be solved up to a common scale factor.

Similarly, let k be an unknown threshold applied to I. Therefore, along level crossings of I we have $k = \sum_j \alpha_j I_j$; hence four level-crossing points that are visible to all four light sources are sufficient for solving α_j and k. $\Box$

The result is that in principle we could cancel the effects of illumination directly from the zero-crossings (or level-crossings) of the novel image instead of from the raw grey-values of the novel image. Note that the model images are represented as before by grey-values (or a continuous transformation of grey-values). Because the model images are taken only once, it is not unreasonable to assume more strict requirements on the quality of those images. We therefore make a distinction between the model acquisition, or learning, phase for which grey-values are used and the recognition phase for which a reduced representation of the novel image is being used.

The result that contours may be used instead of grey-values is not surprising at a theoretical level, considering the literature in image compression. Under certain restrictions on the class of signals, it is known that the zero-crossings form a complete representation of an arbitrary signal of that class. The case of one-dimensional bandpass signals, with certain conditions on the signals' Hilbert transform, is provided by Logan (1977). The more general case is approached by assuming the signal can be represented as a finite complex polynomial (Curtis, Oppenhiem and Lim 1985, Sanz and Huang 1989). Complex polynomials have the well known property that they are fully determined by their *analytic varieties* (curves in the one-dimensional case) using analytic continuation methods (see for example, Saff and Snider 1976). It is well known that analytic continuation is an unstable process (Hille, 1962) and therefore, the reconstruction of the image from its zero-crossings is likely to be unstable. Curtis *et. al.* report, for instance, that zero-crossings must be recorded with great precision, at sub-pixel accuracy of 14 digits.

The result of Proposition 8 can be viewed as a model-based reconstruction theorem, that applies to a much less restricted class of signals (images do not have to be bandpass, for instance). The process is much simpler, but on the other hand it is restricted to a specific model undergoing a restricted group of transformations (changing illumination). The simplicity of the model-based reconstruction, however, is not of great help in circumventing the problem of instability. Stability depends on whether contours are recorded accurately and whether those contours are invariant across the model images.

The assumption that the value of f at a zero-crossing location p is zero, is true for a subpixel location p. In other words, it is unlikely that $f(p) = 0$ for some integral location p. This introduces, therefore, a source of error whose magnitude depends on the 'strength' of the edge that gives rise to the zero-crossing in the signal f, that is, the sharper and stronger the discontinuity in image intensities along an edge in the image I is, the larger the variance around $f(p)$. This suggests that 'weak' edges should be sampled, with more or less the same strength, so that by sampling more than the minimum required number of points, the error could be canceled by a least squares solution.

The second source of error has to do with the stability of the particular edge under changing illumination. Assume, for example, that the zero-crossing at p (recorded accurately) is a result of a sharp change in surface reflectance. Although the image intensity distribution around p changes across the model images, the location of the discontinuity does not, i.e. the zero-crossing is stable. In this case we have that $f(p) = f_j(p) = 0$, $j = 1, 2, 3$. Therefore, such a point will not contribute any information if recorded accurately and will contribute pure noise if recorded with less than the required degree of accuracy. This finding suggests, therefore, that zero-crossings should be sampled along attached shadow contours or along valleys and ridges of image intensities (a valley or a ridge gives rise to two unstable zero-crossings, see Moses 1988).

The situation with reconstruction from level-crossings is slightly different. The first source of error, related to the accuracy in recording the location of level-crossings, still applies, but the second source does not. In general, the variance in intensity around a level crossing point p is not as high as the variance around an edge point. A random sampling of points for a least squares solution is not likely to have a zero mean error, however, and the mean error would therefore be absorbed in the unknown threshold k. The least squares solution would be biased towards a zero mean error solution that will affect both the recovered threshold and the linear coefficients α_j. The solution, therefore, does not necessarily consist of a correct set of coefficients and a slightly off threshold k, but a mixture of both inaccurate coefficients and an inaccurate threshold. This implies that level-crossings should be sampled at locations that do not correspond to zero-crossings in order to minimize the magnitude of errors.

To summarize, the reconstruction of the novel image from three model images and the contours of the novel image is possible in principle. In the case of both zero-crossings and level-crossings, the locations of the contours must be recorded at sub-pixel accuracy. In the case of zero-crossings, another source of potential error arises, which is related to the stability of the zero-crossing location under changing illumination. Therefore, a stable reconstruction requires a sample of points along weak edges that correspond to attached shadow contours or to ridges and valleys of intensity. Alternatively, the locations of contour points must be recorded at sub-pixel accuracy, given also that the sample is large enough to contain unstable points with respect to illumination. Experimental results show that a random sample of ten points (spread evenly all over the object) with accuracy of two digits for zero-crossings and one digit for level-crossings is sufficient to produce results comparable to those produced from sampling image intensities directly. The performance with integral locations of points sampled over edges p that have no corresponding edges in a 3×3 window around p in any of the model images was not satisfactory.

These results show that reconstruction from contours does not appear to be generally useful for the photometric alignment scheme because of its potential instability. It is also important to note that in these experiments the viewing position is fixed, thereby eliminating the correspondence problem that would arise otherwise and would most likely increase the magnitude of errors.

6.2 Photometric Alignment from Contours and Gradients

When zero-crossings are supplemented with gradient data, the reconstruction does no longer suffers from the two sources of errors that were discussed in the previous section. We can can use gradient data to solve for the coefficients, because the operation of taking derivatives (continuous and discrete) is linear and therefore leaves the coefficients unchanged. The accuracy requirement is relaxed because the gradient data is associated with the integral location of contour points, not with their sub-pixel location. Stable zero-crossings do not affect the reconstruction, because the gradient depends on the distribution of grey-values in the neighborhood of the zero-crossing, and the distribution changes with a change in illumination (even though the location of the zero-crossing may not change).

Errors, however, may be more noticeable once we allow changes in viewing positions in addition to changes in illumination (when solving the combined recognition problem). Changes in viewing positions may introduce errors in matching edge points across images. Because the change in image intensity distribution around an edge point is localized and may change significantly at nearby points, then errors in matching edge points across the model images may lead to significant errors in the contribution those points make to the system of equations.

6.3 Photometric Alignment from Sign-bits

Reconstruction from contours, general or model-based, appears to rely on the accurate location of contours. This reliance, however, seems to be at odds with the intuitive interpretation of Mooney-type pictures, like those in Figures 4.3. These images suggest that, instead of contours being the primary vehicle for shape interpretation, the regions bounded by the contours (the sign-bit regions) are primarily responsible for the interpretation process. It is also worthwhile noting that, theoretically speaking, only one bit of information is added in the sign-bit displays. This is because zero-crossings and level-crossings form nested loops (Koenderink and Van Doorn, 1980), and therefore the sign-bit function is completely determined up to a common sign flip. In practice, however, this property of contours does not emerge from edge detectors because weak contours are often thresholded out since they tend to be the most sensitive to noise (see, for example, Figure 4.1). This may also explain why our visual system apparently does not use this property of contours. We therefore do not make use of the global property of the sign-bit function; rather, we treat it as a local source of information, i.e. one bit of information per pixel.

Because the location of contours is an unreliable source of information, especially when the effects of changing viewing positions are considered, we propose to rely instead only on the sign-bit source of information. From a computational standpoint, the only information that a point inside a region can provide is whether the function to be reconstructed (the filtered image f, or the thresholded image I) is positive or negative (or above/below threshold). This information can be incorporated in a scheme for finding a separating hyperplane, as suggested in the following proposition:

Proposition 9 *Solving for the coefficients from the sign–bit image of I is equivalent to solving for a separating hyperplane in 3D or 4D space in which image points serve as "examples".*

Proof: Let $\boldsymbol{z}(p) = (f_1, f_2, f_3)^T$ be a vector function and $\boldsymbol{\omega} = (\alpha_1, \alpha_2, \alpha_3)^T$ be the unknown weight vector. Given the sign-bit filtered image $\hat{f}$ of I, we have that for every point p, excluding zero-crossings, the scalar product $\boldsymbol{\omega}^T \boldsymbol{z}(p)$ is either positive or negative. In this respect, points in $\hat{f}$ can be considered as "examples" in 3D space and the coefficients α_j as a vector normal to the separating hyperplane. Similarly, the reconstruction of the thresholded image $\hat{I}$ can be represented as a separating hyperplane problem in 4D space, in which $\boldsymbol{z}(p) = (I_1, I_2, I_3, -1)^T$ and $\boldsymbol{\omega} = (\alpha_1, \alpha_2, \alpha_3, k)^T$. $\square$

The contours lead to a linear system of equations, whereas the sign-bits lead to a linear system of *inequalities*. The solution to a linear system of inequalities $A\boldsymbol{w} < \boldsymbol{b}$ can be approached using Linear Programming techniques or using Linear Discriminant Analysis techniques (see Duda and Hart 1973 for a review). Geometrically, the unknown weight vector $\boldsymbol{w}$ can be considered as the normal direction to a plane, passing through the origin, in 3D Euclidean space, and a solution is found in such a way that the plane separates the "positive" examples, $\boldsymbol{\omega}^T \boldsymbol{z}(p) > 0$, from the "negative" examples, $\boldsymbol{\omega}^T \boldsymbol{z}(p) < 0$. In the general case, where $\boldsymbol{b} \neq 0$, the solution is a point inside a polytope whose faces are planes in 3D space.

The most straightforward solution is known as the *perceptron* algorithm (Rosenblatt, 1962). The basic perceptron scheme proceeds by iteratively modifying the estimate of $\boldsymbol{w}$ by the following rule:

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n + \sum_{i \in M} \boldsymbol{z}^i$$

where $\boldsymbol{w}^n$ is the current estimate of $\boldsymbol{w}$, and M is the set of examples $\boldsymbol{z}^i$ that are incorrectly classified by $\boldsymbol{w}^n$. The critical feature of this scheme that it is guaranteed to converge to a solution, irrespective of the initial guess $\boldsymbol{w}^0$, provided that a solution exists (examples are linearly separable). Another well known method is to reformulate the problem as a least squares optimization problem of the form

$$\min_{\boldsymbol{w}} |A\boldsymbol{w} - \boldsymbol{b}|^2$$

where the i'th row of A is $\boldsymbol{z}^i$, and $\boldsymbol{b}$ is a vector of arbitrarily specified positive constants (often $\boldsymbol{b} = \mathbf{1}$). The solution $\boldsymbol{w}$ can be found using the pseudoinverse of A, i.e.

$$\boldsymbol{w} = A^{+}\boldsymbol{b} = (A^t A)^{-1} A^t \boldsymbol{b},$$

or iteratively through a gradient descent procedure, which is known as the Widrow-Hoff procedure. The least squares formulation is not guaranteed to find a correct solution but has the advantage of finding a solution even when a correct solution does not exist (a perceptron algorithm is not guaranteed to converge in that case).

By using the sign-bits instead of the contours, we are trading a unique, but unstable, solution for an approximate, but stable, solution. The stability of reconstruction from sign-bits is achieved by sampling points that are relatively far away from the contours. This sampling process also has the advantage of tolerating a certain degree of misalignment between the images as a result of less than perfect correspondence due to changes in viewing position (this feature is discussed further in Chapter 8). Experimental results (see Figures 6.1 and 6.2) demonstrate that 10 to 20 points, distributed over the entire object, are sufficient to produce results that are comparable to those obtained from an exact solution. The experiments were done on images of 'Ken' and on another set of face images taken from a plaster bust of Roy Lamson (courtesy of the M.I.T Media Laboratory). We tried both the perceptron algorithm and the least-squares approach and found that both yielded practically the same results. The sample points were chosen manually, and over several trials we found that the reconstruction is not sensitive to the particular choice of sample points, as long as they are not clustered in a local area of the image and are sampled a few pixels away from the contours. The results (see Figures 6.1 and 6.2) show the reconstruction of a novel thresholded images from three model images. The linear coefficients and the threshold are recovered from the system of inequalities using a sample of 16 points; the model images are then combined and thresholded with the recovered threshold to produce a synthesized thresholded image. Recognition then proceeds by matching the novel thresholded image given as input against the synthesized image.

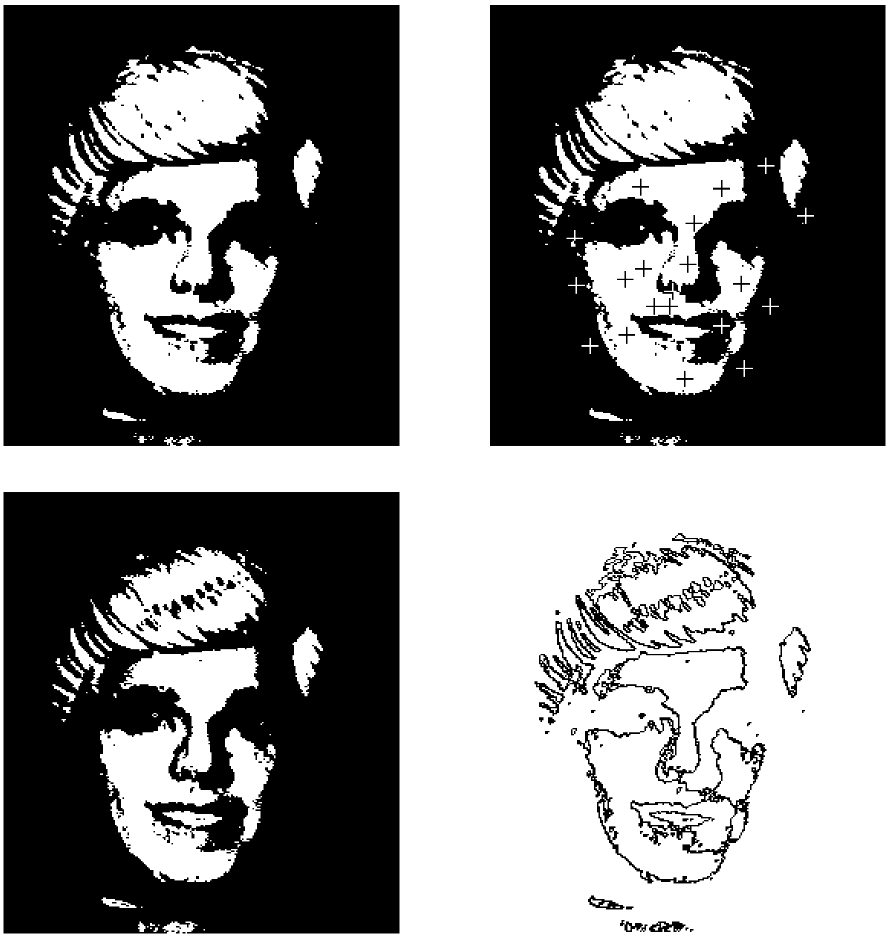


Figure 6.1: Reconstruction from sign-bits. *Top Row (left to right):* the input novel image; the same image but with the sample points marked for display. *Bottom Row:* the reconstructed image; the overlay of the original level-crossings and the level-crossings of the reconstructed thresholded image.

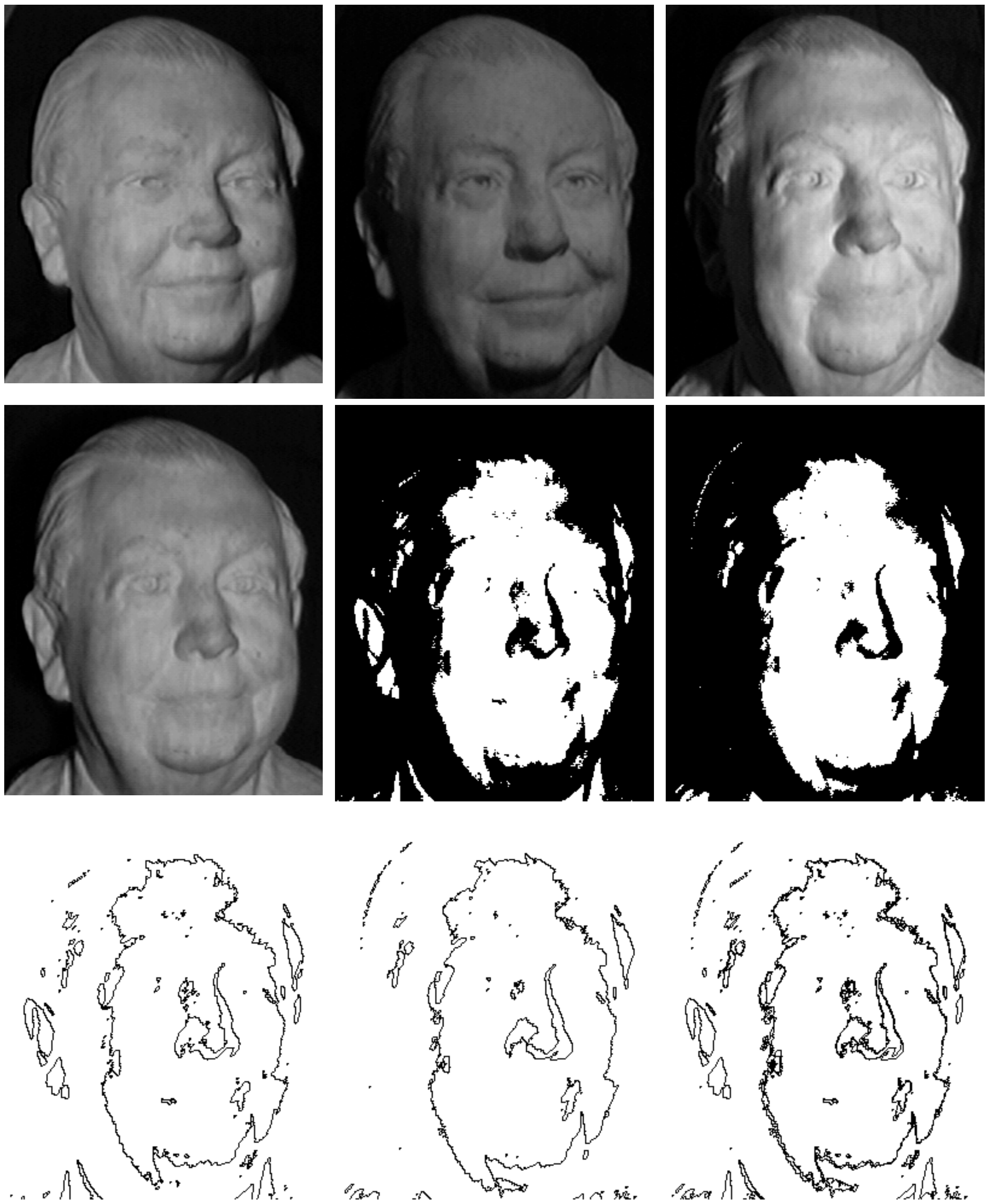


Figure 6.2: Reconstruction from sign-bits. *Row 1:* three model images. *Row 2:* novel image; thresholded input; reconstructed image (same procedure as described in the previous figure). Note that the left ear has not been reconstructed; this is mainly because the ear is occluded in two of the three model images. *Row 3:* the level-crossings of the novel input; level-crossings of the reconstructed image; the overlay of both level-crossing images.

6.4 Summary of Part II

In this part of the thesis we addressed the problem of recognition under changing illumination conditions. Unlike the geometric problem of recognition, the photometric problem has not received much attention in the past and, therefore, we devoted Chapter 4 for motivating and exploring this problem by use of examples drawn from empirical observations on human vision and from computational schemes in related areas of visual analysis. We have arrived at two conclusions. First, there appears to be a need for a model-based approach to the photometric problem. Second, the process responsible for factoring out the illumination during the recognition process appears to require more than contour information, but just slightly more.

We have seen that a possible model-based method for dealing with illumination changes is photometric stereo. In this method multiple images of the same object taken from different illumination conditions are recorded and are then used to recover scene information. We have seen that the major problem with photometric stereo is that one must either assume that illumination parameters are known a priori or instead assume that the surface albedo is uniform across the surface. We suggested as an alternative using a method, we call photometric alignment, that is also based on recording multiple images of the object. We do not use these images in order to recover intrinsic properties of the object, as used in photometric stereo, but rather to directly compensate for the change in illumination conditions for any other novel image of the object. This difference is critical, for it enables us to avoid the limitations of photometric stereo by allowing an arbitrary distribution of surface albedo and by not having to assume or recover the parameters associated with the light sources.

Assuming that the photometric alignment scheme is the process responsible for factoring out the illumination during the recognition process, our objective in this chapter was to explore the possibilities of using less than image grey-values for this purpose. Specifically, we were interested in making a computational connection with the empirical observation made in Chapter 4 that sign-bits appear to be sufficient for visual interpretation, whereas edges alone do not. The connection was made by introducing two new results: first, step edges and level-crossings of the novel image are theoretically sufficient for the photometric alignment scheme. This result, however, assumes that edges be given at sub-pixel accuracy — a finding that implies difficulties in making use of this result in practice. Second, the sign-bit information can be used instead of edges.

Photometric alignment using sign-bits is a region-based process by which points inside the binary regions of the sign-bit image are sampled and each contributes a partial observation. Taken together, the partial observations are sufficient to determine the solution for compensating for illumination. The more points sampled, the more accurate the solution. Experimental results show that a relatively small number of points (10 to 20) are generally sufficient for obtaining solutions that are comparable to those obtained by using the image grey-levels. This method agrees with the empirical observations that were made in Chapter 4 regarding the possibility of having a region-based process rather than a contour-based one, the possibility of preferring sign-bits over edges, and the sufficiency of sign-bits for factoring out the illumination. Finally, the possibility of using sign-bits instead of edges raises a potentially practical issue related to changing viewing positions. A region-based computation has the advantage of tolerating a small degree of misalignment between the images due to changing viewing positions. This finding implies that the illumination can be factored out even in the presence of small changes in viewing positions without explicitly addressing the geometric problem of compensating for viewing transformations. We discuss this property further in Chapter 8.

Part III

Geometry and Photometry: Correspondence and the Combined Recognition Problem

The Problem of Achieving Full Correspondence

Chapter 7

In this chapter we address the problem of achieving full correspondence between the model views. This problem arises during the model acquisition stage of representing the object by a small number of images. In the geometric domain the model images were taken from different viewing positions, and in the photometric domain those images were taken from different illumination conditions. In the generl case, we must deal with the problem of achieving correspondence between all interest points across the model images which are taken under different viewing positions and different illumination conditions. Achieving full correspondence is a critical component of the overall scheme of combining geometric and photometric sources of variabilities for purposes of recognition.

In Part I we distinguished between two kinds of correspondences: minimal correspondence and full correspondence. Minimal correspondence involves matching a small number of points (four, six, or eight) between the novel image and the model images in order to recover the alignment transformation. This matching is assumed to take place during recognition based on a small number of distinct features which, presumably, can be detected regardless of changing viewing positions and illumination conditions. Full correspondence involves the matching of all points of interest across the model images. Note that the phrase all interest points actually means all image points across the model views because re-projection is to be achieved at both the geometric and photometric levels. Full correspondence is assumed to take place during the model acquisition stage, rather than during recognition. Unlike the problem of minimal correspondence, however, we cannot simply assume that points can be matched across images unaffected by changing viewing positions and changing illumination.

We approach the problem of achieving full correspondence between two images of an object in the following manner: first, we assume that minimal correspondence is available between the two images. Second, we assume that the images are taken under similar illumination conditions. The first assumption enables us to apply the geometric results described in Part I, that is, recovering the epipolar geometry between the two views. The second assumption enables the use of changing grey-levels between the two images in order to solve for correspondence wherever the image gradients are not vanishing and are not perpendicular to the epipolar line direction. We will see in the next chapter that this basic approach is sufficient for achieving full correspondence between the model views of the combined recognition problem, i.e., in the situation where the model images are taken from different viewing positions and different illumination conditions.

Similar to the analogy between the geometric problem of recognition and the problem of structure from motion, there is a strong connection between the problem of achieving full correspondence and the problem of visual motion (the analogy between the two is discussed in more detail in the next section). We will, therefore, use terms taken from from the area of visual motion — such as optical flow or dense flow — interchangeably with full correspondence throughout this chapter.

7.1 Correspondence and Optical Flow: Brief Review

The general problem of achieving correspondence or optical flow, is to recover the two-dimensional displacement field between points in both images. The problem is generally difficult and various approaches have been proposed in the literature. The difficulty arises primarily because the displacement field depends on the three-dimensional structure of the scene and the particular viewing geometry or motion of the camera, niether of which are known in advance.

One generally distinguishes between attempts to recover a sparse and discrete type of correspondence and attempts to recover a dense and often continuous type of correspondence. The discrete correspondence methods generally aim at establishing a discrete point-to-point match between a sparse set of points in both images. The methods of solution to this type of problem tend to focus less on the geometrical aspects of 3D to 2D in terms of viewing geometry and projections, and more on the combinatorial aspect of searching for a best match under various optimization constraints for reducing the search, such as uniqueness, continuity along curves, order constraint, measures of affinity (Ullman 1979, Marr and Poggio 1979, Thompson and Barnard 1981, Grimson 1982, Hildreth 1984, Baird 1985).

The dense correspondence methods often assume small, or infinitesimal motion, in which case the displacement field is a velocity field. The methods of solution to this type of problem tend to rely entirely on the instantaneous spatio-temporal patterns of image grey-values, and are often referred to as optical flow methods. Optical flow techniques can be divided into three major classes: (i) differential techniques, (ii) region-based matching techniques, and (iii) energy-based techniques. Differential techniques rely on the instantaneous spatial and temporal derivatives of image intensity in order to determine the velocity vector up to an unknown component in the direction perpendicular to the intensity gradient vector. This assumes that the change in image intensity is due entirely to the motion of the camera or the scene, and not to photometric effects, such as changing direction of light sources. The remaining component of the velocity vector is determined by using some form of smoothness constraint, or by introducing higher order derivatives at the expense of restricting further the admissible velocity field (Horn and Schunk 1981, Lucas and Kanade 1981, Glazer *et. al.* 1983, Verri and Poggio 1989, Nagel 1987).

Using cross-correlations or sum of squares difference (SSD) measures of matching quality, region-based techniques of optical flow attempt to find the best match between image regions in one view and neighboring regions in the other view (Lucas 1984, Anandan 1987). Energy-based methods rely on the response of velocity-tuned filters, such as oriented Gabor filters or Reichardt detectors (Adelson and Bergen 1985, Van Santen and Sperling 1985, Heeger 1987).

The methods for achieving optical flow share a fundamental limitation known as the aperture problem: the spatio-temporal pattern of intensity can provide only one component of the velocity vector. The remaining component can be recovered provided we assume that velocity does not change across the region of inspection and, in addition, that the region contains sufficient intensity structure (sufficient amount of variation in gradient direction across the region, which often occurs at corners, or high curvature, of intensity).

The correspondence methods (discrete and continuous) described so far do not make significant use of the geometrical constraints that follow from having two projections of a three-dimensional scene. Waxman and Wohn (1985) and Bachelder and Ullman (1992) suggest methods for correspondence that account for the 3D to 2D geometry in a way that is limited to locally planar surfaces. Waxman and Wohn suggest an approach by which the surface is broken down into local planar patches, and they derive correspondence using the observation that planar surfaces under perspective projection give rise to a quadratic flow field (Waxman and Ullman, 1985). As with the method of Waxman and Ullman, the smaller the patch size the more unstable the system becomes because of narrowing of the field of view (see Adiv, 1989). Bachelder and Ullman (1992) suggest a method for measuring correspondence along curves, using orthographic projection; they also assume local planarity. The difference is that planarity is assumed along curves, rather than over patches, which has the advantage that the plane is not restricted to being tangent to the surface, thereby locations that require a large support for reliably measuring correspondence may still satisfy the planarity assumption, even though the surface is not planar.

The general idea behind the approach presented in this chapter is to put together the source of information coming from the spatio-temporal pattern of image intensity (as in optical-flow techniques) and the geometric source of information that arises from assuming a rigid world projected onto the image plane. The geometrical source of information can be captured by having a small number of corresponding points between the two images. Another way to view this approach is that a small number of correspondences are sufficient for recovering correspondences everywhere else. Minimal correspondence can be found using standard optical flow techniques that are applied over regions associated with surface markings (see for instance Anandan 1987, Tomasi 1991, for automatically detecting such regions).

7.2 Correspondence from two Views Under Parallel Projection

Consider again the equations 2.1 and 2.2 relating the affine coordinates b_1, b_2, b_3 of an object point P (with respect to a basis O, P_1, P_2, P_3 of four other non-coplanar object points) with the corresponding points p and p' in two views of the object created by means of parallel projection. These equations are reproduced below:

$$op = \sum_{j=1}^{3} b_j(op_j)$$
$$o'p' = \sum_{j=1}^{3} b_j(o'p'_j).$$

These equations were introduced in Section 2.2 for purposes of recovering the affine coordinates of P, given that we have all the correspondences we need. We can also view these equations from the standpoint of obtaining the location of p', given the correspondences due to the four reference points and the affine coordinates of P. Since we do not have a sufficient number of observations to recover the affine coordinates, we need to look for an additional source of information.

We assume that the correspondences due to the four reference points are known, that is, we have solved for minimal correspondence, and that both views are taken under similar illumination conditions:

$$I(x + \Delta x, y + \Delta y, t + 1) = I(x, y, t),$$

where $\boldsymbol{v} = (\Delta x, \Delta y)$ is the displacement vector, i.e., $p' = p + \boldsymbol{v}$. We assume the convention that the two views were taken at times t and $t+1$. A first order approximation of a Taylor series expansion leads to the following equation which describes a linear approximation to the change of image grey-values at p due to motion:

$$\nabla I \cdot \boldsymbol{v} + I_t = 0, \tag{7.1}$$

where ∇I is the gradient at point p, and I_t is the temporal derivative at p. Equation 7.1 is known as the "constant brightness equation" and was introduced by Horn and Schunk (1981). In addition to assuming that the change in grey-values is due entirely to motion, we have assumed that the motion (or the size of view separation) is small, and that the surface patch at P is locally smooth. In practice, the size of view separation can be traded off with the smoothness of the surface by using coarse-to-fine techniques — as described later in this chapter.

The constant brightness equation provides only one component of the displacement vector v, the component along the gradient direction, or normal to the isobrightness contour at p. This "normal flow" information is sufficient to uniquely determine the affine coordinates b_j at p, as shown next. By subtracting equation 2.1 from equation 2.2 we get the following relation:

$$\boldsymbol{v} = \sum_{j=1}^{3} b_j \boldsymbol{v}_j + (1 - \sum_j b_j)\boldsymbol{v}_o, \tag{7.2}$$

where $\boldsymbol{v}_j$ $(j = 0, .., 3)$ are the known displacement vectors of the points o, p_1, p_2, p_3. By substituting equation 7.2 in the constant brightness equation, we get a new equation in which the affine coordinates are the only unknowns:

$$\sum_j b_j [\nabla I \cdot (v_j - v_o)] + I_t + \nabla I \cdot v_o = 0. \tag{7.3}$$

Equations 2.1, and 7.3, provide a complete set of linear equations to solve for the affine coordinates at all locations p that have a non-vanishing gradient, which is not perpendicular to the direction of the epipolar line passing through p'. Once the affine coordinates are recovered, the location of p' immediately follows. We have, therefore, arrived to the following result:

Proposition 10 (4pt + brightness) *Two parallel projected images of a shaded 3D surface with four clearly marked reference points admit a complete set of linear equations representing the affine coordinates of all surface points, provided that the surface is undergoing an infinitesimal affine transformation and that the two images are taken under identical illumination conditions.*

In practice, it is more convenient to recover p' using the affine structure representation, rather than using affine coordinates. The derivation of affine structure in Section 2.3 concluded with,

$$o'p' = A(op) + \gamma_p \boldsymbol{w},$$

where $\boldsymbol{w} = o'p'_3 - A(op_3)$ is the epipolar line direction, A is the matrix $[o'p'_1, o'p'_2][op_1, op_2]^{-1}$, and γ_p is an invariant measure of a deviation of P from the reference plane passing through O, P_1, P_2. The unknown parameter γ_p can be recovered from the equation,

$$\gamma_p = \frac{-I_t - \nabla I \cdot [A - I](op)}{\nabla I \cdot \boldsymbol{w}}.$$

A convenient way to view this result is that the location of the corresponding point $o'p'$ is determined by a "nominal motion", described by $A(op)$ and a "residual motion", described by $\gamma_p \boldsymbol{w}$. The nominal motion component is determined only from the minimal correspondence information (the correspondence due to the four reference points), and the residual motion component is determined with the help of the constant brightness equation 7.1.

There are two reasons for considering the overall displacement as composed of two components, nominal and residual. First, from a practical point of view we would like to handle situations of long range motion (relatively wide view separation) between the two views, and therefore, limit as much as possible the contribution of the constant brightness equation. Because γ_p is a measure of "affine depth", the smaller the depth variation between the surface and the reference plane, the smaller the residual motion component becomes (assuming rigid motion and approximately orthographic projection). Therefore, with surfaces that do not extend much in depth we can achieve longer ranges of motion by first compensating for the nominal motion and then recovering the residual motion component. This process is described in more detail in Section 7.4. The second reason is more speculative in nature: the separation of overall motion into two components suggests that the measurement of motion is conducted relative to a frame of reference. The frame of reference is determined by the motion of a small number of key points, and these, in turn, provide a first approximation for motion everywhere else within that frame. The approximation is accurate if the moving object is a plane, otherwise it is refined by solving for the residual motion component. In the next section we attempt to draw a connection with empirical observations of human vision. This connection may support the existence of this kind of a two-stage computation in the measurement of visual motion.

7.2.1 Frame of Reference and the Measurement of Motion

The notion of a frame of reference that precedes the computation of motion may have some support, albeit indirectly, in human vision literature. The phenomenon of "motion capture" introduced by Ramachandran (Ramachandran 1986, Ramachandran and Cavanagh 1985, Ramachandran and Inada 1985) is suggestive to the kind of motion measurement presented here. Ramachandran and his collaborators observed that the motion of certain salient image features (such as gratings or illusory squares) tend to dominate the perceived motion in the enclosed area by masking incoherent motion signals derived from uncorrelated random dot patterns, in a winner-take-all fashion. Ramachandran therefore suggested that motion is computed by using salient features that are matched unambiguously and that the visual system assumes that the incoherent signals have moved together with those salient features. The scheme suggested in this chapter may be viewed as a refinement of this idea. Motion is "captured" in Ramachandran's sense for the case of a planar surface in motion, not by assuming the motion of the the salient features, but by computing the nominal motion transformation. For a non-planar surface the nominal motion is only a first approximation which is further refined by use of spatio-temporal detectors, provided that the remaining residual displacement is in their range, namely, the surface captured by the frame of reference is sufficiently flat. In this view the effect of capture attenuates with increasing depth of points from the reference plane, and is not affected, in principle, by the proximity of points to the salient features in the image plane.

The motion capture phenomenon also suggests that the salient features that are selected for providing a frame of reference must be spatially arranged to provide sufficient cues that the enclosed pattern is indeed part of the same surface. In other words, not any arrangement of four non-coplanar points, although theoretically sufficient, is an appropriate candidate for a frame of reference. This point has also been raised by Subirana-Vilanova and Richards (1991) in addressing perceptual organization issues. They claim that convex image chunks are used as a frame of reference that is imposed in the image prior to constructing an object description for recognition. The frame then determines inside/outside, top/bottom, extraction/contraction and near/far relations that are used for matching image constructs to a model.

Other suggestive data include stereoscopic interpolation experiments by Mitchison and McKee (1985). They describe a stereogram with a central periodic region bounded by unambiguously matched edges. Under certain conditions the edges impose one of the expected discrete matchings (similar to stereoscopic capture, see also Prazdny 1986). In other conditions a linear interpolation in depth occurred between the edges violating any possible point-to-point match between the periodic regions. The linear interpolation in depth corresponds to a plane passing through the unambiguously matched points. This observation may support the idea that correspondence starts with the computation of nominal motion, which is determined by a small number of salient unambiguously matched points, and is later refined using short-range motion mechanisms.

7.3 Correspondence under a Wide Field of View

The assumption of parallel projection holds approximately for settings in which the object occupies a relatively narrow field of view. One way to extend this method to wider fields of view is to assume central projection instead. Under central projection a similar correspondence method would require eight corresponding points as a minimal correspondence requirement, rather than four. The details are apparent once we consider the correspondence method as proceeding by first recovering the epipolar geometry (with which we can determine correspondence up to an unknown location along the epipolar line passing through p') followed by the use of the constant brightness equation to determine the location of p' along its epipolar line. Section 3.8 provides the details of recovering the epipolar geometry from eight points under central projection.

Another approach is to apply the correspondence method locally by making use of the geometric interpretation of having a reference plane and a reference point for the nominal transformation. Given that we have a set of $n > 4$ corresponding points, we can form a triangulation on the set of points. The triangulation divides the image into regions, each with three corresponding points, within which the correspondence method can be applied independently of other regions (the fourth point can be taken from a neighboring triangle). Because all neighboring triangles share an edge, the particular solution for triangulation does not affect the resulting flow field (Huttenlocher and Ullman 1987 used triangulation for extending the three-point alignment method to non-rigid objects).

7.4 Implementation Using a Coarse-to-fine Architecture

The use of the constant brightness equation for determining the residual motion term αw assumes that $|\gamma_p \boldsymbol{w}|$ is small. In practice, the residual motion is not sufficiently small everywhere and, therefore, a hierarchical motion estimation framework is adopted for the implementation. The assumption of small residual motion is relative to the spatial neighborhood and to the temporal delay between frames; it is the ratio of the spatial to the temporal sampling step that is required to be small. Therefore, the smoother the surface the larger the residual motion that can be accommodated. The Laplacian Pyramid (Burt and Adelson, 1983) is used for hierarchical estimation by refining the estimation of γ_p at multiple resolutions. The rationale being that large residuals at the resolution of the original image are represented as small residuals at coarser resolutions, therefore satisfying the requirement of small displacement. The γ_p estimates from previous resolutions are used to bring the image pair into closer registration at the next finer resolution.

The particular details of implementation follow the "warp" motion framework suggested by Lucas and Kanade (1981), Bergen and Adelson (1987) and by Bergen and Hingorani (1990). Described in a nutshell, a synthesized intermediate image is first created by applying the nominal transformation to the first view. To avoid subpixel coordinates, we actually compute flow from the second view towards the first view. In other words, the intermediate frame at location p contains a bilinear interpolation of the brightness values of the four nearest pixels to the location $\tilde{p}' = A(op) + o'$ in the first view, where the 2D affine transformation A is computed from view 2 to view 1. The γ field is estimated incrementally by projecting previous estimates at a coarse resolution to a finer resolution level. Gaps in the estimation of γ_p, because of vanishing image gradients or other low confidence criteria, are filled-in at each level of resolution by means of membrane interpolation. Once the γ field is projected to the finer level, the displacement field is computed (the vector $\gamma_p \boldsymbol{w}$) and the two images, the intermediate and the second image, are brought into closer registration. This procedure proceeds incrementally until the finest resolution has been reached.

7.4.1 Experimental Results

The correspondence method was applied to images of 'Ken' undergoing rigid rotation, mainly around the vertical axis. Four snapshots were taken covering a range of about 23 degrees of rotation. The light setting consisted of two point light sources located in front of the object, 60 degrees apart.

Minimal correspondence of four points was obtained from the flow field generated by the warp motion algorithm (Lucas and Kanade 1981, Bergen and Adelson 1987, Bergen and Hingorani 1990) along points having good contrast at high spatial frequencies, e.g., the tip of the eyes, mouth and eye-brows (those points were selected manually, but in principle they can be determined automatically using measures of confidence, such as in Anandan 1987, Tomasi 1991).

The combination of the particular light setting and the complexity of the object make it a challenging experiment for two reasons: (i) the object is sufficiently complex to have cast shadows and specular points, both of which undergo a different motion than the object itself, and (ii) because of the light setting, the change in grey-values across the views is not due entirely to motion but also due to change in relative illumination conditions between the object and the light sources.

The results of correspondence in all these experiments are displayed in several forms. The flow field is displayed to illustrate the stability of the algorithm, indicated by the smoothness of the flow field. The first image is 'warped', i.e., all image points are displaced by the amount specified by the computed flow to create a synthetic image that should match the second image. The warped image is displayed in order to check for deformations (or lack there of). Finally, the warped image is compared with the second image by superimposing, or taking the difference of, their edge images that were produced using a Canny (1983) edge detector with the same parameter settings.

7.4.2 Incremental Long Range Motion

In this experiment, flow was computed independently between each consecutive pair of images, using a fixed set of four reference points, and then combined to form a flow from the first image, Ken1, to the fourth image, Ken4. The rationale behind this experiment is that because shape is an integral part of computing correspondence/flow, then flow from one consecutive pair to the next should add up in a consistent manner.

Figure 7.1 shows the results on the first pair of images, Ken1 and Ken2, separated by 6^o rotation. As expected, the location of strong cast shadows (one near the dividing hair line) and specular points in the warped image do not match those in Ken2. The superimposed edge images illustrate that correspondence is accurate, at least up to a pixel accuracy level. The flow field is smooth even in the case where no explicit smoothing was done.

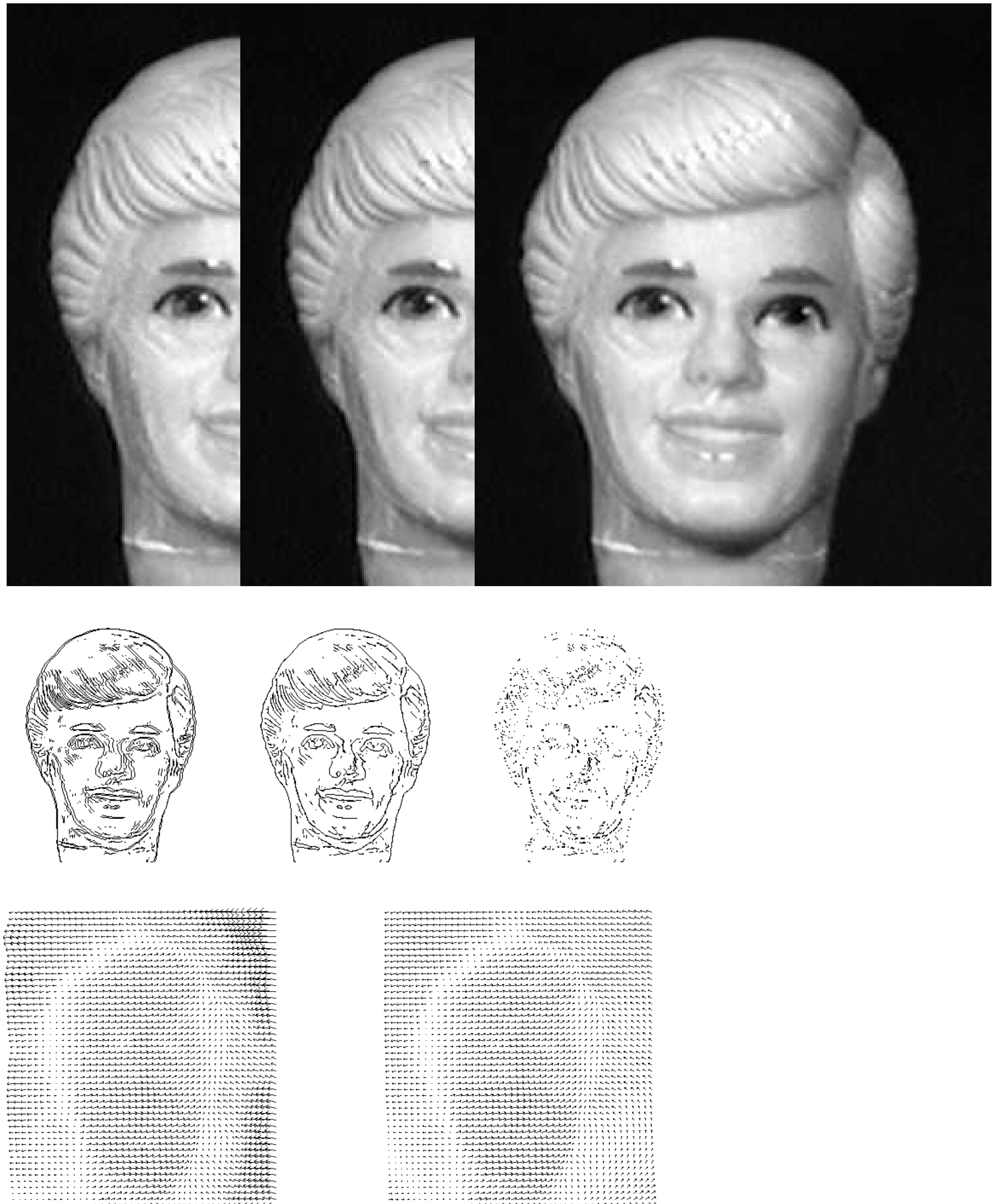


Figure 7.1: Results of shape and correspondence for the pair Ken1 and Ken2. *First (top) row:* Ken1,Ken2 and the warped image Ken1-2. *Second row:* Edges of Ken1 and Ken2 superimposed, edges of Ken2 and Ken1-2 superimposed, difference between edges of Ken2 and Ken1-2. *Third row:* Flow field in the case where γ_p is estimated in a least squares manner in a 5×5 sliding window, and flow field when γ_p is computed at a single point (no smoothing).

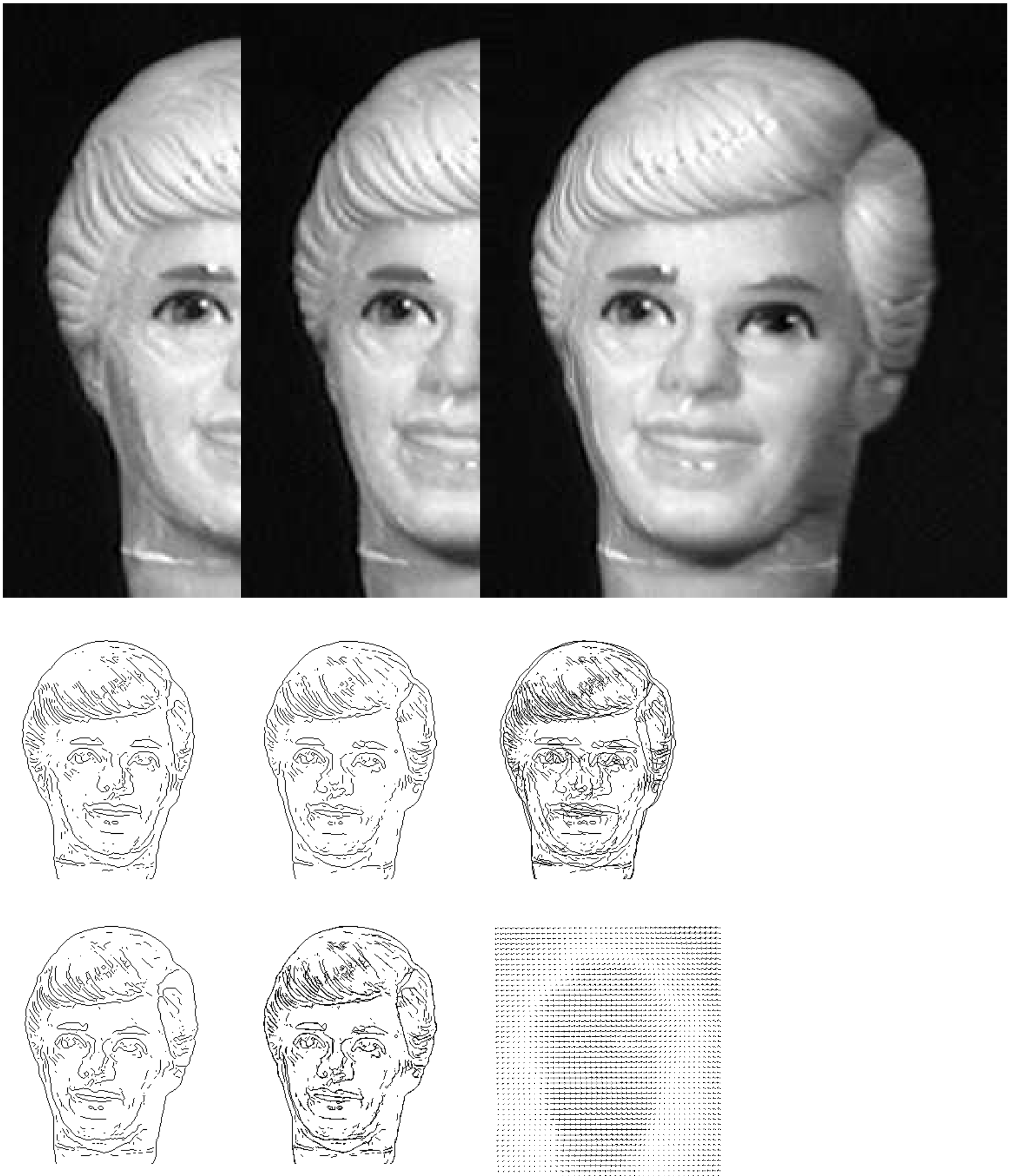


Figure 7.2: Results of combining flow from Ken1 to Ken4. First row: Ken1,Ken4 and the warped image Ken1-4. Second row: edges of Ken1, Ken4 and edges of both superimposed. Third row: edges of Ken1-4, edges of Ken4 and edges of Ken1-4 superimposed, flow field from Ken1 to Ken4 (scaled for display).

Figure 7.2 shows the results of combining flow across consecutive pairs computed independently (using the same four reference points) to produce flow from Ken1 to Ken4. Except for the point specularities and the strong shadow at the hair line, the difference between the warped image and Ken4 is only at the level of difference in brightness (because of change in viewing angle). No apparent deformation is observed in the warped image. The flow field is as smooth as the flow from Ken1 to Ken2, implying that flow was combined in a consistent manner.

7.4.3 Comparison With Optical Flow Methods

With the correspondence method presented here we have exchanged the smoothness assumption, used in optical flow methods, with the assumption that the world is rigid (undergoing parallel or central projection). The rigidity-based approach is, therefore, less general than smoothness-based optical flow methods. The question we address in this section is whether there is a practical reason for preferring the rigidity-based method over other, more general, optical flow methods. In many practical situations full correspondence is being sought for purposes of recovering rigid structure from motion, or for purposes of modeling a rigid structure by full correspondence between two or more of its views. The images of 'Ken', for example, are particularly challenging for smoothness-based methods because of the relative small number of intensity corners in the image. As a result, a relatively small number of "good" correspondences would determine, by means of smoothness, the correspondences everywhere else.

We applied two well-known optical flow methods: a differential technique following Lucas and Kanade (1981) and Adelson and Bergen (1987), and a region-based technique due to Anandan (1987). Both algorithms received good reviews in a recent quantitative study held by Barron, Fleet, Beauchemin and Burkitt (1991). We used the implementation of Anandan's method found in KB-Vision (image processing shell) written at the University of Massachusetts. The implementation of the Lucas-Kanade technique was adopted from Bergen and Hingorani (1990).

Figure 7.3 displays the resulting flow field produced by both algorithms on the pair Ken1 and Ken2 (short-range motion). The quality of the flow field (in subjective terms of smoothness and regularity) is slightly better with Lucas-Kanade's algorithm, than Anandan's. The flow field produced by our method looks smoother and has fewer flow vectors that change direction in an abrupt manner. We applied next Lucas-Kanade's algorithm to the sequence Ken1 to Ken4 to see how stable the flow field is when flow is combined

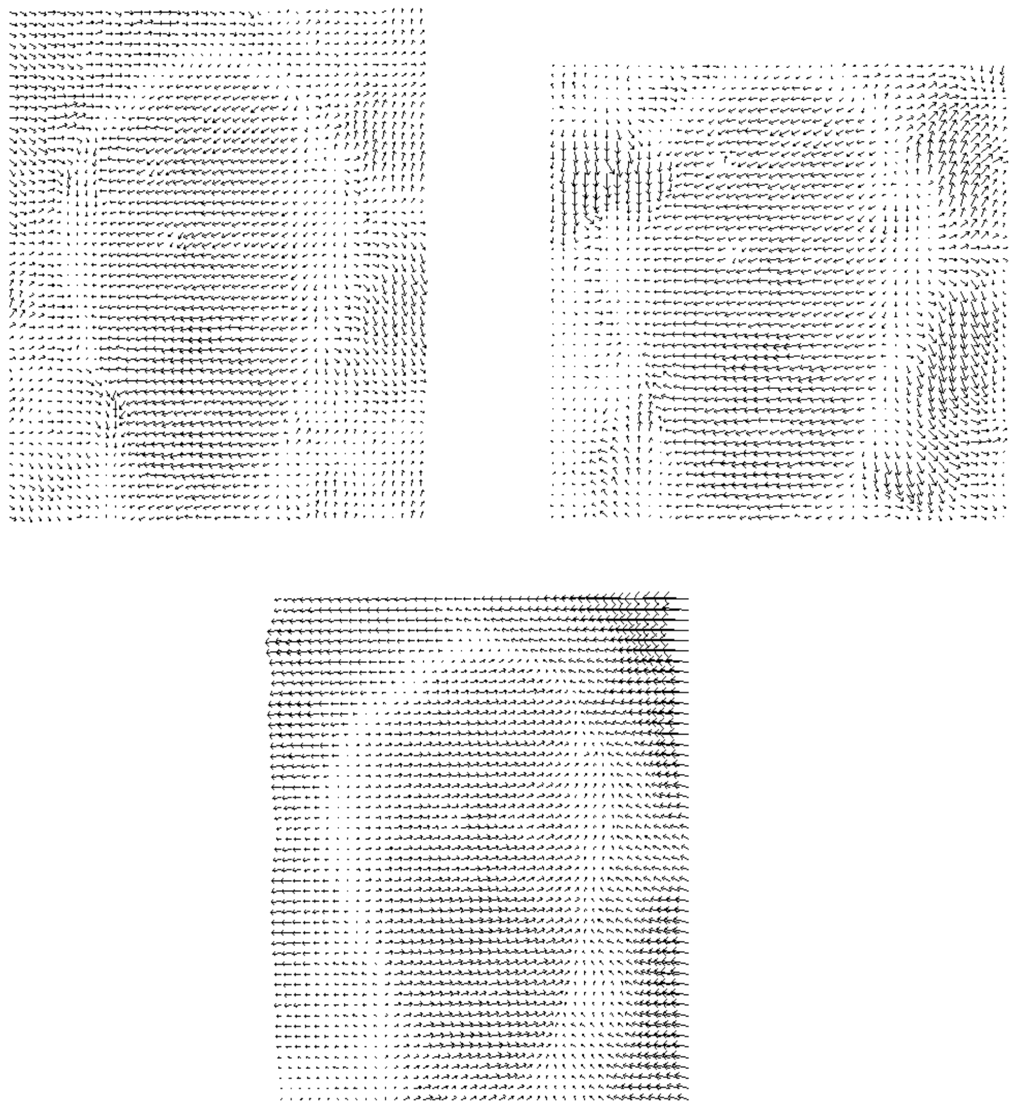


Figure 7.3: Flow field on the pair Ken1 and Ken2, produced by alternative methods. *Row 1:* left image is the flow produced by Lucas-Kanade algorithm, right image is the flow produced by Anandan's algorithm. *Row 2:* The flow produced by our algorithm (for comparison).

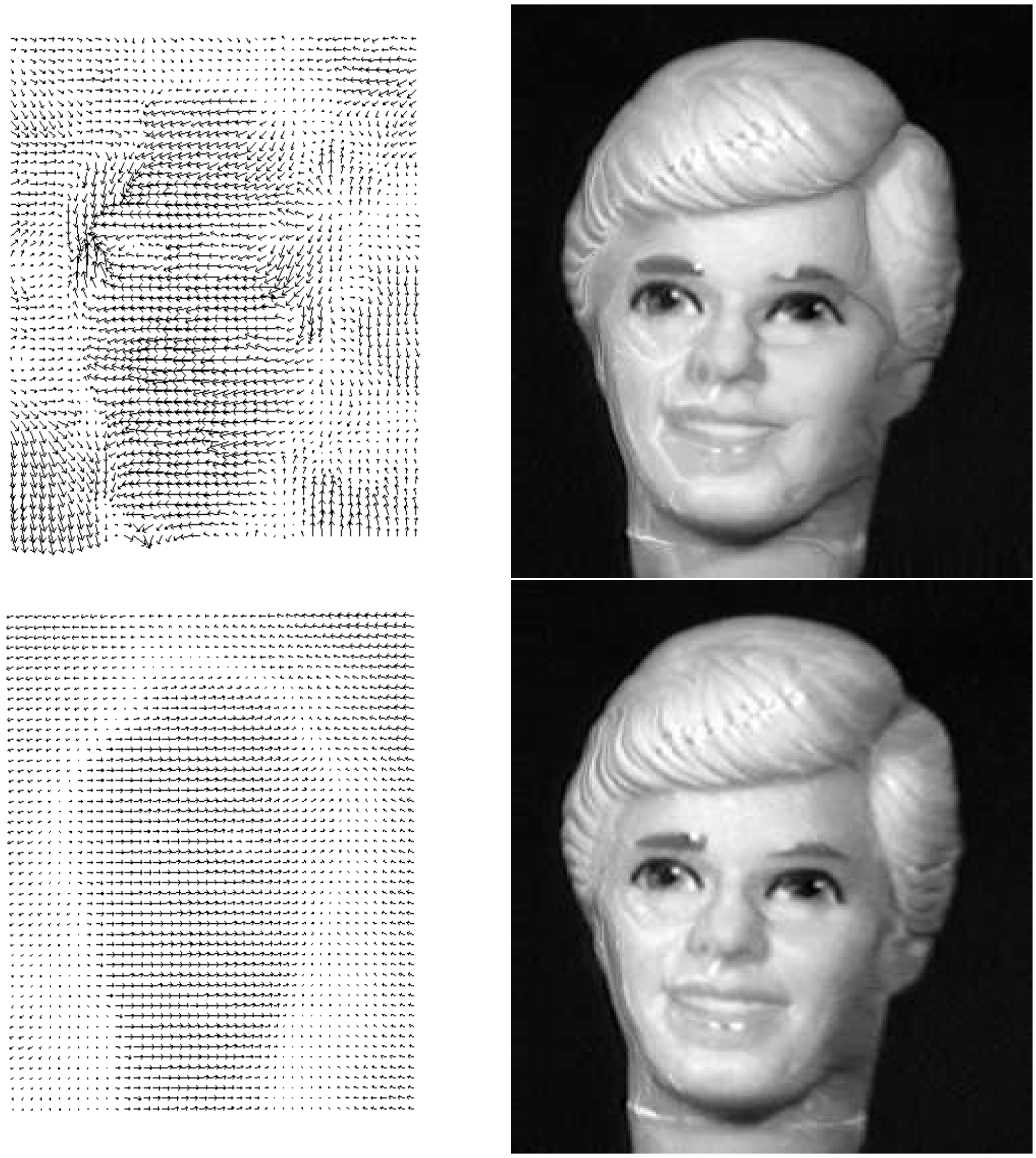


Figure 7.4: Flow field incrementally added from Ken1 to Ken4. *Top Row:* left image is the flow produced by Lucas-Kanade algorithm, right image is the warped image created by warping Ken1 using the computed flow. *Bottom Row:* The flow and warped image produced by our algorithm (for comparison).

incrementally along each consecutive pair. If the flow between each consecutive pair is not sufficiently accurate, then the incremental addition will cause errors to accumulate and produce an overall distorted flow field. As we saw in Figure 7.2, this did not happen with our algorithm and the flow was added in a consistent manner. Figure 7.4 shows the results for the Lucas-Kanade algorithm. The flow field is considerably less smooth than before and has several distinct pockets of discontinuity in the direction of flow vectors. Also shown is the 'warped' image using the Lucas-Kanade flow field where the lack of smoothness is apparent in several distinct deformations in the warped image.

7.4.4 Long Range Motion

The two-stage scheme for measuring motion — nominal motion followed by a short-range residual motion detection — suggests that long-range motion can be handled in an area enclosed by the privileged points. The restriction of short-range motion is replaced by the restriction of limited depth variation from the reference plane. As long as the depth variation is limited, then correspondence should be obtained regardless of the range of motion. Note that this is true as long as we are sufficiently far away from the object's bounding contour. The larger the rotational component of motion — the larger the number of points that go in and out of view. Therefore, we should not expect good correspondence at the boundary. The claim that is tested in the following experiment, is that under long range motion, correspondence is accurate in the region enclosed by the frame of reference, i.e., points that are relatively far away from the boundary.

Figure 7.5 shows the results of computing flow directly from Ken1 to Ken4. Note the effect of the nominal motion transformation applied to Ken1. The nominal motion brings points closer together inside the frame of reference; points near the boundary are taken farther apart from their corresponding points because of the large depth difference between the object's rim and the reference plane. The warped image looks very similar to Ken4 except near the boundary. The deformation at the boundary may be due to both the relatively large residual displacement, remaining after nominal motion was applied, and to the repetitive intensity structure of the hair; Therefore it may be that the frequency of the hair structure caused a misalignment at some level of the pyramid which was propagated.

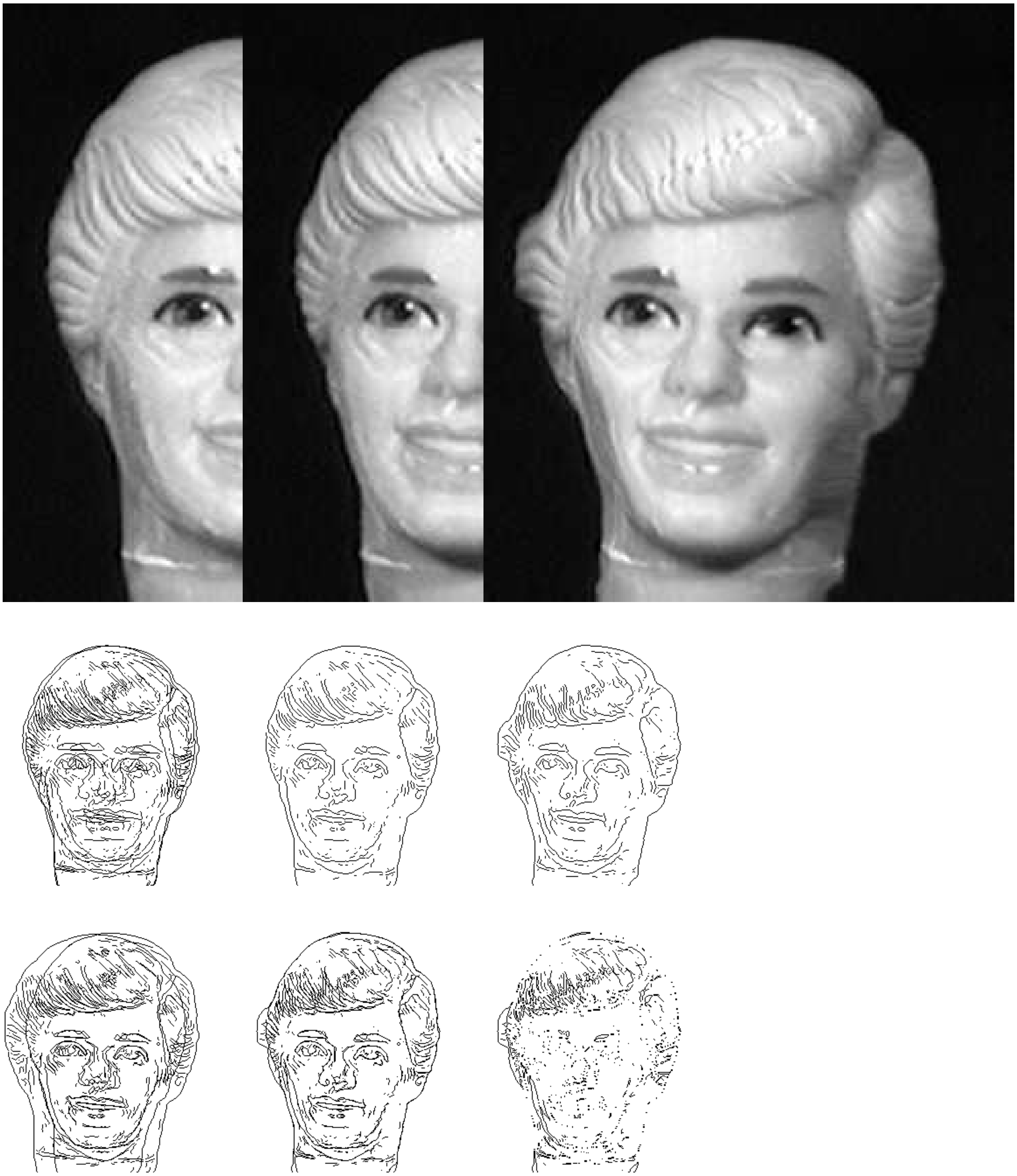


Figure 7.5: Results of computing long-range flow from Ken1 to Ken4. *First (top) row:* Ken1,Ken4 and the warped image Ken1-4. *Second row:* Edges of Ken1 and Ken4 superimposed, edges of Ken4 and edges of Ken1-4. *Third row:* Edges of Ken4 superimposed on edges of the nominal transformed Ken1, edges of Ken4 and Ken1-4 superimposed, and difference between edges of Ken4 and edges of Ken1-4.

7.5 Chapter Summary

We have approached the correspondence problem by combining two sources of information into a single computational scheme. One source of information comes from assuming that the object is rigid, and therefore a small number of known correspondences can constrain correspondences everywhere else, and the second source of information comes from the spatio-temporal image intensity distribution. Taken together, these sources of information completes the system of equations for determining correspondence for all other points. The full correspondence problem is, therefore, reduced to the problem of achieving minimal correspondence, i.e., finding a small number of corresponding points whose detection is unaffected by geometric and photometric transformations.

The Combined Recognition Problem: Geometry and Illumination

Chapter 8

We have described so far three components that are necessary building blocks for dealing with recognition via alignment under the geometric and photometric sources of variability. First, is the component describing the geometric relation between two model views and a novel view of an object of interest. Second, is the component describing the photometric relation between three model images and a novel image of the object. Third, is the correspondence component with which it becomes possible to represent objects by a small number of model images. The geometric and photometric components were treated independently of each other. In other words, the photometric problem assumed the surface is viewed from a fixed viewing position. The geometric problem assumed that the views are taken under a fixed illumination condition, i.e., the displacement of feature points across the different views is due entirely to a change of viewing position. In practice, the visual system must confront both sources of variability at the same time. The combined geometric and photometric problem was defined in Section 1.2 and is reproduced below:

Combined Problem: *We assume we are given three model images of a 3D matte object taken under different viewing positions and illumination conditions. For any input image, determine whether the image can be produced by the object from some viewing position and by some illumination condition.*

The combined problem definition suggests that the problem be solved in two stages: first, changes in viewing positions are compensated for, such that the three model images are aligned with the novel input image. Second, changes of illumination are subsequently compensated for, by using the photometric alignment method. In the following sections we describe several experiments with 'Ken' images starting from the procedure that was used for creating the model images, followed by three recognition situations: (i) the novel input image is represented by its grey-levels, (ii) the input representation consists of sign-bits, and (iii) the input representation consists of grey-levels, but the model images are taken from a fixed viewing position (different from the viewing position of the novel image). In this case we make use of the sign-bits in order to achieve photometric alignment although the novel image is taken from a different viewing position.

8.1 Creating a Model of the Object

The combined recognition problem implies that the model images represent both sources of variability, i.e., be taken from at least two distinct viewing positions and from three distinct illumination conditions. The three model images displayed in the top row of Figure 8.1 were taken under three distinct illumination conditions, and from two distinct viewing positions (23^o apart, mainly around the vertical axis). In order to apply the correspondence method described in the previous chapter, we took an additional image in the following way. Let the three illumination conditions be denoted by the symbols S_1, S_2, S_3, and the two viewing positions be denoted by V_1, V_2. The three model images, from left to right, can be described by $< V_1, S_1 >, < V_2, S_2 >$ and $< V_1, S_3 >$, respectively. Since the first and third model images are taken from the same viewing position, the two images are already aligned. In order to achieve full correspondence between the first two model images, a fourth image $< V_2, S_1 >$ was taken. Correspondence between $< V_1, S_1 >$ and $< V_2, S_1 >$ was achieved via the correspondence method described in the previous chapter. Since $< V_2, S_1 >$ and $< V_2, S_2 >$ are from the same viewing position, then the correspondence achieved previously holds also between the first and second model images. The fourth image $< V_2, S_1 >$ was then discarded and did not participate in subsequent recognition experiments.

8.2 Recognition from Grey-Level Images

The method for achieving recognition under both sources of variability is divided into two stages: first, the three model images are re-projected onto the novel image. This is achieved by solving for minimal correspondence between the novel image and one of the model images. With minimal correspondence of four points across the images (model and novel) we can predict the new locations of model points that should match with the novel image (in central projection we need six or eight points). Second, photometric alignment is subsequently applied by selecting a number of points (no correspondence is needed at this stage because all images are now view-compensated) to solve for the linear coefficients. The three model images are then linearly combined to produce a synthetic image that is both view and illumination compensated, i.e., should match the novel image.

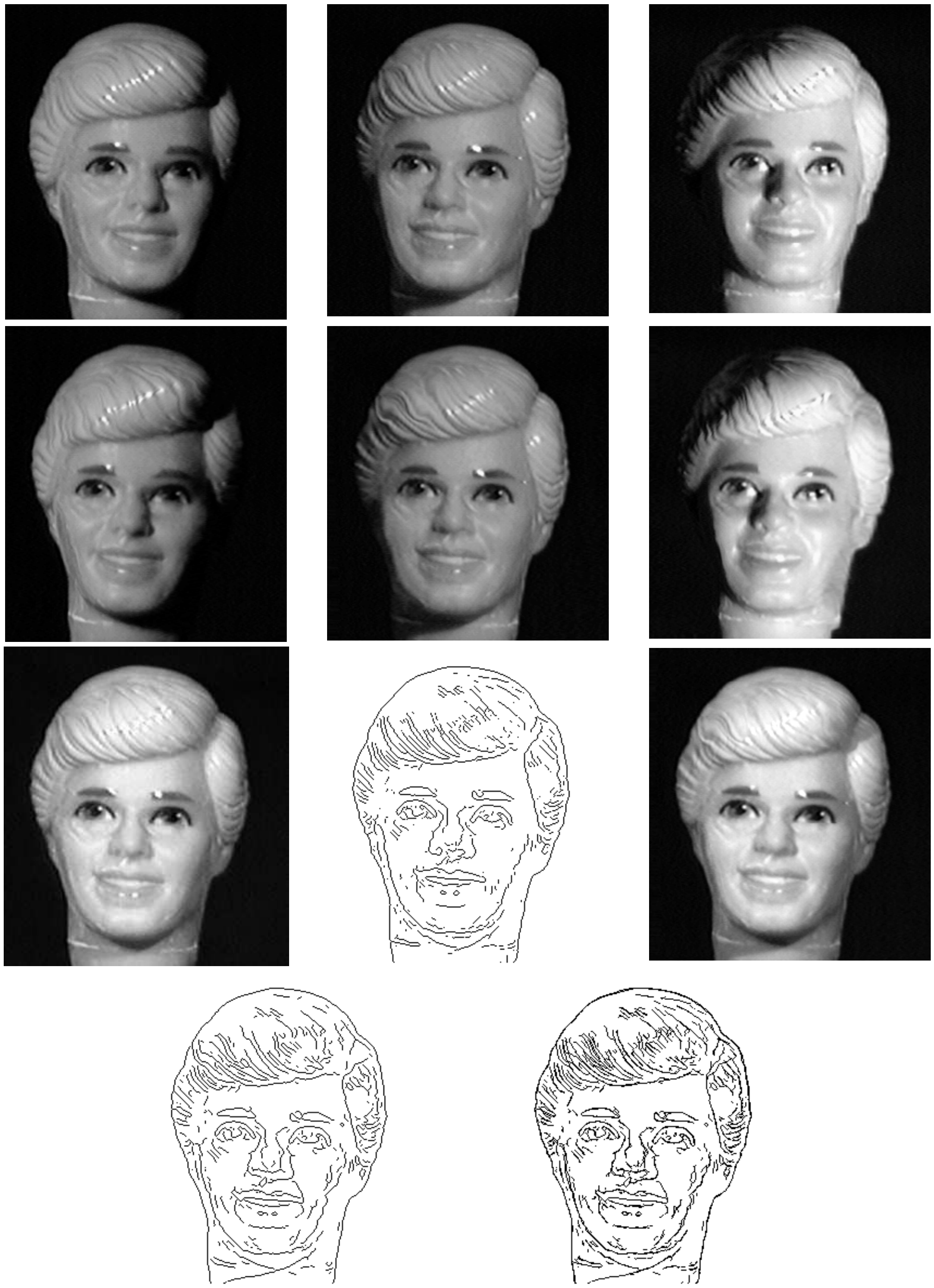


Figure 8.1: Recognition from full grey-level novel image (see text for more detailed description). *Row 1 (left to right):* Three model images (the novel image is shown third row lefthand display). *Row 2:* View-compensated model images — all three model images are transformed (using four points) as if viewed from the novel viewing position. *Row 3:* Novel image, edges of novel image, photometric alignment of the three view-compensated model images (both view and illumination compensated). *Row 4:* Edges of the resulting synthesized image (third row righthand), overlay of edges of novel and synthesized image.

Figure 8.1 illustrates the chain of alignment transformations. The novel image, displayed in the third row left image, is taken from an in-between viewing position and illumination condition. Although, in principle, the recognition components are not limited to in-between situations, there are few practical limitations. The more extrapolated the viewing position is, the more new object points appear and old object points disappear, and similarly, the more extrapolated the illumination condition is, the more new cast shadows are created (see Section 5.1.1). Minimal correspondence was achieved by manually selecting four points that corresponded to the far corners of the eyes, one eye-brow corner, and one mouth corner. These points were matched across the model views by applying the warp motion algorithm (Lucas and Kanade 1981, Bergen and Adelson 1987). Re-projection was then achieved by using the affine structure method, described in Section 2.3 (which is the same as applying the linear combination of views method). Then the model views were re-projected onto the novel view, and their original grey-values retained. As a result, we have created three synthesized model images (shown in Figure 8.1, second row) that are from the same viewing position as the novel image, but have different image intensity distributions due to changing illumination. The photometric alignment method was then applied to the three synthesized model images and the novel image, without having to deal with correspondence because all four images were already aligned. The sample points for the photometric alignment method were chosen automatically by searching over smooth regions of image intensity (as described in Section 5.1.3). The resulting synthesized image is displayed in Figure 8.1, third row right image. The similarity between the novel and the synthesized image is illustrated by superimposing the step edges of the two images (Figure 8.1, bottom row right image).

Almost identical results were obtained by assuming central projection and using the 6-point scheme. Two additional points for the minimal correspondence were selected: an eye-brow corner, and the other mouth corner (the two eye corners and the two mouth corners are approximately coplanar). The 8-point method requires in practice slightly more corresponding points (10 points in a least-squares solution for the epipoles were sufficient for achieving comparable results), which was partly due to the fact that this particular object does not contain many points that can be reliably matched using optical flow techniques (i.e., points at corners of intensity).

Since 'Ken' images in this experiment are approximately orthographic, the remaining experiments were done under the assumption of parallel projection, i.e., we used either the affine structure method or the linear combination of views. It is worthwhile noting, that with parallel projection, the combined recognition problem can be solved by simply applying certain linear combinations of the model views.

8.3 Recognition from Reduced Images

A similar procedure to the one described above can be applied to recognize a reduced novel image. In this case the input image is taken from a novel viewing position and illumination condition followed by a thresholding operator (unknown to the recognition system). Figure 8.2 illustrates the procedure. We applied the linear combination method of re-projection and used more than the minimum required four points. In this case it is more difficult to extract corresponding points between the thresholded input and the model images reliably. Therefore, seven points were manually selected and their corresponding points were manually estimated in the model images. The linear combination method was then applied using a least squares solution for the linear coefficients to produce three synthesized view-compensated model images. The photometric alignment method from sign-bits was then applied (Section 6.3) using a similar distribution of sample points as shown in Figure 6.1.

We consider next another case of recognition from reduced images, in which we make use of the property exact alignment is nor required when using sign-bits.

8.4 Recognition from a Single Viewing Position

Photometric alignment from sign-bits raises the possibility of compensating for changing illumination without needing an exact correspondence between the model images and the novel image. The reason lies in the way points are sampled for setting the system of inequalities, that is, points are sampled relatively far away from the contours (see Section 6.3). In addition, the separation of image displacements into nominal and residual components (Section 7.2) suggests that in an area of interest bounded by at least three reference points, the nominal transformation alone may be sufficient to bring the model images close enough to the novel image so that we can apply the photometric alignment from sign bits method.

Consider, for example, the effect of applying only the nominal transformation between two different views (Figure 8.3). Superimposing the two views demonstrates that the displacement is concentrated mostly in the center area of the face (most likely the area in which we would like to select the sample points). By selecting three corresponding points

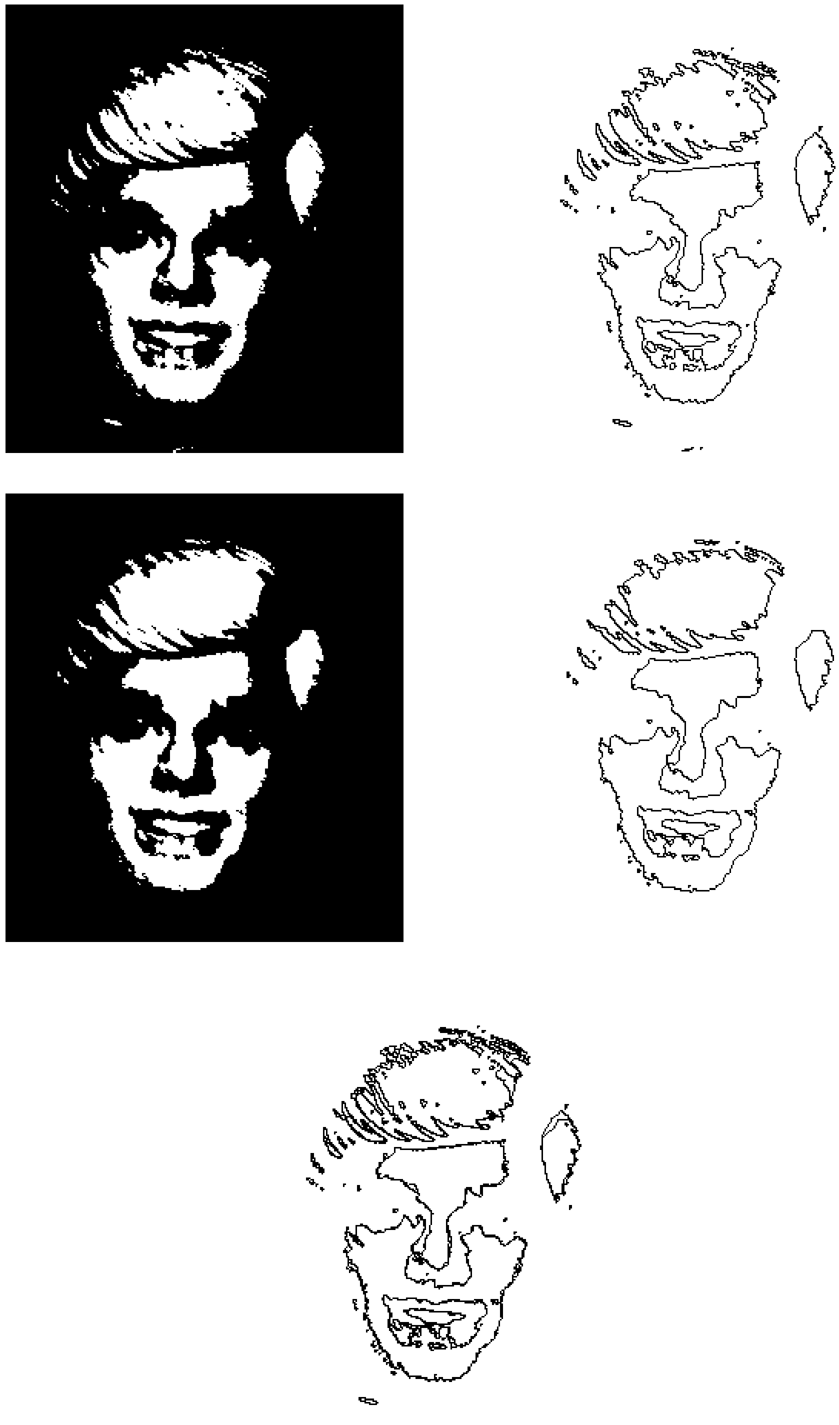


Figure 8.2: Recognition from a reduced image. *Row 1 (left to right):* novel thresholded image; its level-crossings (the original grey-levels of the novel image are shown in the previous figure, third row on the left). *Row 2:* the synthesized image produced by the recognition procedure; its level-crossings. *Row 3:* overlay of both level-crossings for purposes of verifying the match.

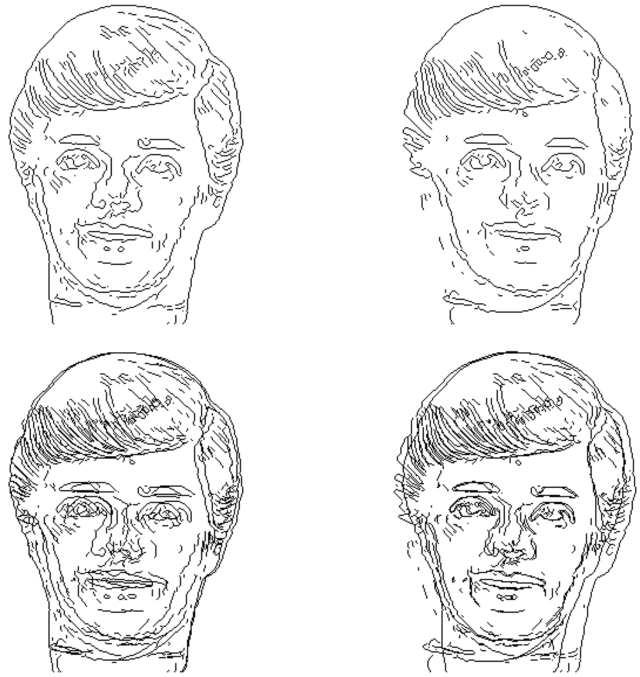


Figure 8.3: Demonstrating the effect of applying only the nominal transformation between two distinct views. *Row 1:* edges of two distinct views. *Row 2:* overlay of both edge image, and overlay of the edges of the left image above and the nominally transformed righthand image.

covering the center area of the face (two extreme eye corners and one mouth corner), the 2D affine transformation (nominal transformation) accounts for most of the displacement in the area of interest at the expense of large displacements at the boundaries (Figure 8.3, bottom row on the right).

Taken together, the use of sign-bits and the nominal transformation suggests that one can compensate for illumination and for relatively small changes in viewing positions from model images taken from the same viewing position. We apply first the nominal transformation to all three model images and obtain three synthesized images. We then apply the photometric alignment from sign-bits to recover the linear coefficients used for compensating for illumination. The three synthesized images are then linearly combined to obtain an illumination-compensated image. The remaining displacement between the synthesized image and the novel image can be recovered by applying the residual motion transformation (along the epipolar direction using the constant brightness equation).

Figure 8.4 illustrates the alignment steps. The three model images are displayed in the top row and are the same as those used in Chapter 5 for compensating for illumination alone. The novel image (second row, left display) is the same as in Figure 8.1, i.e., it is taken from a novel viewing position and novel illumination condition. The image in the center of the second row illustrates the result of attempting to recover the correspondence (using the full correspondence method described in the previous chapter) between the novel image and one of the model images without first compensating for illumination. The image on the left in the third row is the result of first applying the nominal transformation to the three model images followed by the photometric alignment using the sign-bits (the sample points used by the photometric alignment method are displayed in the image on the right in the second row). The remaining residual displacement between the latter image and the novel image is recovered using the full correspondence method and the result is displayed in the center image in the third row. The similarity between the final synthesized image and the novel image is illustrated by superimposing their step edges (fourth row, right display).

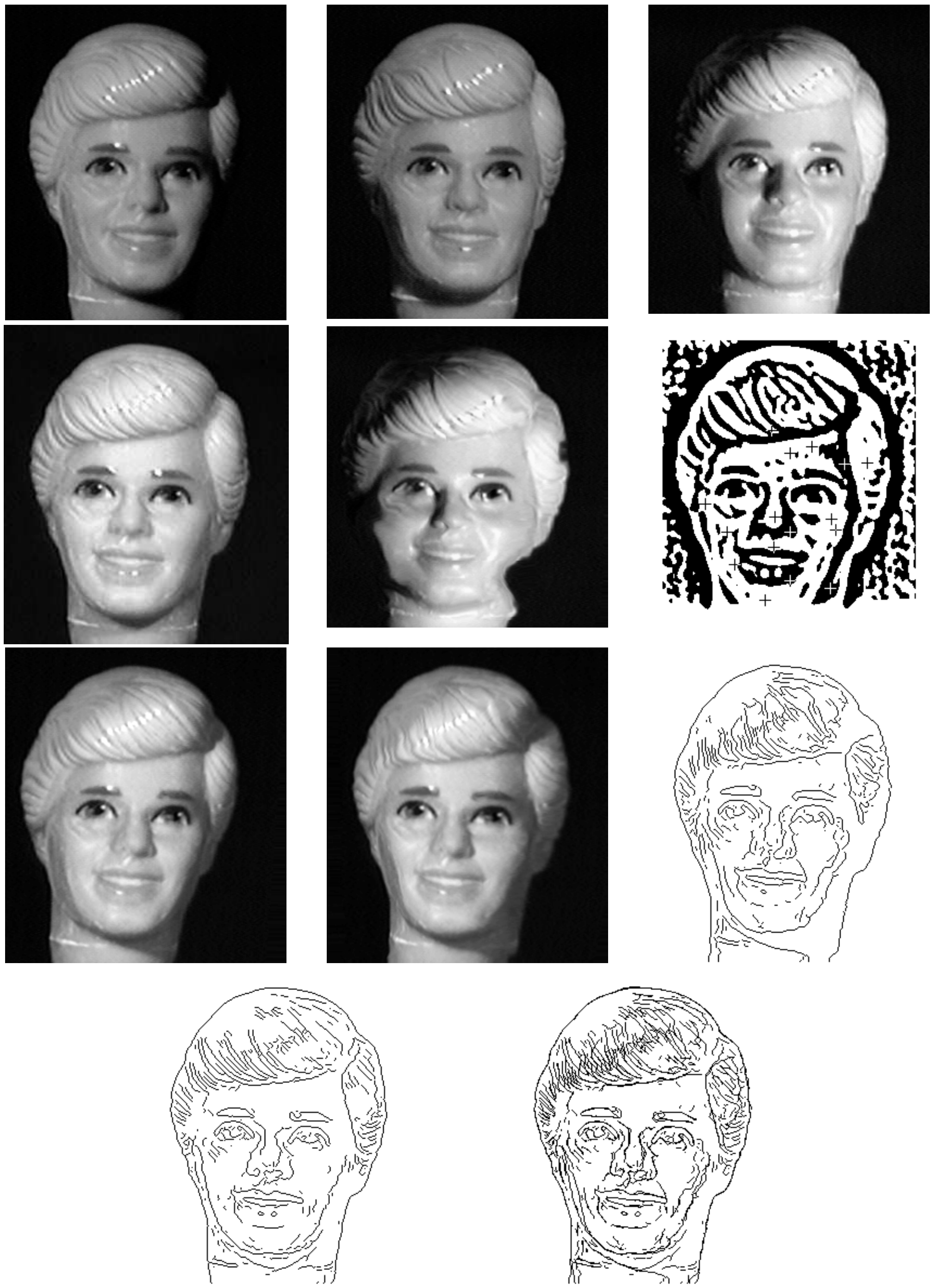


Figure 8.4: Recognition from a single viewing position (see text for details).

Conclusions and Discussion

Chapter 9

This chapter provides an opportunity to step back and look over the range of issues and technical results presented in this thesis. Our major goal was to gain new understandings of geometric and photometric issues relevant to visual recognition. The starting point of our research was that computer visual recognition has almost exclusively focused on only one source of variability, i.e., the geometric problem of recognition under changing viewing positions. Moreover, the methods for handling the geometric problem, both in recognition as well as in SFM, leaves open many important issues. For example, we have argued that the transition between orthographic and perspective models is problematic, the notion of camera calibration is also problematic, and the representation of structure is largely an open issue that has not received much attention, yet has significant ramifications on the kind of technical results that can be obtained.

The new technical results presented in this thesis are largely based on the way we viewed the range of geometric and photometric issues. In the geometric domain we emphasized three central issues: first and foremost, is that the transition from parallel to central projection can be made natural and transparent if we have the same representation of structure under both projection models. This, for example, implied that previous work on affine structure should be extended in the way of introducing a new geometric invariant, rather than to recover projective coordinates. Second, is the use of non-metric methods for recognition and SFM by means of adopting the model of central projection. Third, is the connection between alignment-based recognition and SFM. We have emphasized the similarity between the two by showing that in parallel projection one can derive an alignment scheme (the linear combination of views) that appears not to involve structure nor camera geometry from a SFM method (affine structure from two views). This implies that although our main interest is to achieve recognition via alignment, it may be useful to approach the problem from the standpoint of SFM.

In the photometric domain we observed, both from a practical point of view and from empirical observations of human vision, that changing illumination is a source of variability that in some cases appears to be factored out during the recognition process in a model-based manner. Related to that we have also observed that edges alone are sometimes not sufficient for visual interpretation, but slightly more than edges are sufficient. The central question is then, what information from the image should be carried on to high level visual processes, and how is this information used within the context of recognition? This direction is substantially different from the mainstream approach of treating the photometric aspect mostly at the level of feature or edge detection.

The technical contributions made in this thesis can be divided into three parts: geometric related contributions, photometric related, and contributions related to the combination of both sources of information.

- The major technical contribution in the geometric part was made in Theorem 1 by showing that besides recovering the epipoles, parallel projection and central projection are essentially the same. In other words, a relative structure invariant, that holds under both projection models, can be defined relative to four points in space and, moreover, it can be uniquely recovered from two views regardless of whether one or the other was created by means of parallel or central projection.

- The technical contributions in the photometric part included the photometric alignment method, and the use of sign-bit information for achieving recognition.

- The method for achieving full correspondence between two views provided a technical contribution in the domain of putting together geometry (assumption of rigidity) and grey-values into a single computational scheme. This approach differs from the mainstream of current methods for achieving correspondence or optical flow. Instead of assuming an arbitrary smooth transformation between the two images, we assumed that the two images are different projections (parallel or central) of the same rigid object. This assumption together with the spatio-temporal image intensity distribution is sufficient for obtaining correspondence.

Finally, we have shown how the geometric, photometric, and the correspondence components can be put together to solve for the combined recognition problem, i.e., recognition of an image of a familiar object taken from a novel viewing position and a novel illumination condition.

9.1 Future Directions

The photometric part of this thesis has been developed to a lesser extent than the geometric part. The reason is partly due to the lack of prior research in this domain, and partly due to the relatively large number of related issues that did not fall within the scope of this thesis. We sketch below some of these issues.

The ability to interpret Mooney images of faces may suggest that these images are an extreme case of a wider phenomenon. Some see it as a tribute to the human ability to separate shadow borders from object borders (Cavanagh, 1990); in this thesis we have noted that the phenomenon may indicate that in some cases illumination is factored out in a model-based manner and that the process responsible apparently requires more than just contour information, but only slightly more. A possible topic of future research in this domain would be to draw a connection, both at the psychophysical and computational levels, between Mooney images and more natural kinds of inputs. For example, images seen in newspapers, images taken under poor lighting, and other low quality imagery have less shading information to rely on and their edge information may be highly unreliable, yet are interpreted without much difficulty by the human visual system. Another related example, is the image information contained in draftsmen's drawings. Artists rarely use just contours in their drawings and rely on techniques such as "double stroking" to create a sense of relief (surface recedes towards the contours) and highlights to make the surface protude. These pictorial additions that artists introduce are generally not interpretable at the level of contours alone, yet do not introduce any direct shading information.

Another related topic of future interest is the level at which sources of variability are compensated for. In this thesis the geometric and photometric sources of variability were factored out based on connections between different images of individual objects. The empirical observations we used to support the argument that illumination should be compensated for in a model-based manner, actually indicate that if indeed such a process exists, it is likely to take place at the level of classifying the image as belonging to a general class of objects, rather than at the level of identifying the individual object. This is simply because the Mooney images are of generally unfamiliar faces, and therefore, the only model-based information available is that we are looking at an image of a face. A similar situation may exist in the geometric domain as well, as it is known that humans can recognize novel views just from a single view of the object.

There are also questions of narrower scope related to the photometric domain that may be of general interest. The question of image representation in this thesis was applied only to the novel image. A more general question should apply to the model acquisition stage as well. In other words, what information needs to be extracted from the model images, at the time of model acquisition, in order to later compensate for photometric effects? This question applies to both the psychophysical and computational aspects of the problem. For example, can we learn to generalize to novel images just from observing many Mooney-type images of the object? (changing illumination, viewing positions, threshold, and so forth). A more basic question is whether the Mooney phenomenon is limited exclusively to faces. And if not, what level of familiarity with the object, or class of objects, is necessary in order to generalize to other Mooney-type images of the same object, or class of objects.

At a more technical level, there may be interest in further pursuing the use of sign-bits. The sign-bits were used as a source of partial observations that, taken together, can restrict sufficiently well the space of possible solutions for the photometric alignment scheme. In order to make further use of this idea, and perhaps apply it to other domains, the question of how to select sample points, and the number and distribution of sample points, should be addressed in a more systematic manner.

Finally, regarding the connection between projective structure and alignment under central projection. We have shown that in parallel projection the linear combination of views can be derived from the method of recovering affine structure from two views. In order to close the loop, it may be of interest to show a similar connection in central projection and as a result extend the linear combination result to one that applies to central projection. We know that this is possible and plan to do it in the future.

Fundamental Theorem of Plane Projectivity

Appendix A

The fundamental theorem of plane projectivity states that a projective transformation of the plane is completely determined by four corresponding points. We prove the theorem by first using a geometric drawing, and then algebraically by introducing the concept of rays (homogeneous coordinates). The appendix ends with the system of linear equations for determining the correspondence of all points in the plane, given four corresponding points (used repeatedly throughout this paper).

Definitions: A *perspectivity* between two planes is defined as a central projection from one plane onto the other. A *projectivity* is defined as made out of a finite sequence of perspectivities. A projectivity, when represented in an algebraic form, is called a *projective transformation.* The fundamental theorem states that a projectivity is completely determined by four corresponding points.

Geometric Illustration

Consider the geometric drawing in Figure A.1. Let A, B, C, U be four coplanar points in the scene, and let A', B', C', U' be their projection in the first view, and A'', B'', C'', U'' be their projection in the second view. By construction, the two views are projectively related to each other. We further assume that no three of the points are collinear (four points form a quadrangle), and without loss of generality let U be located within the triangle ABC. Let BC be the x-axis and BA be the y-axis. The projection of U onto the x-axis, denoted by U_x, is the intersection of the line AU with the x-axis. Similarly U_y is the intersection of the line CU with the y-axis. because straight lines project onto straight lines, we have that U_x, U_y correspond to U'_x, U'_y if and only if U corresponds to U'. For any other point P, coplanar with $ABCU$ in space, its coordinates P_x, P_y are constructed in a similar manner. We therefore have that B, U_x, P_x, C are collinear and therefore the cross ratio must be

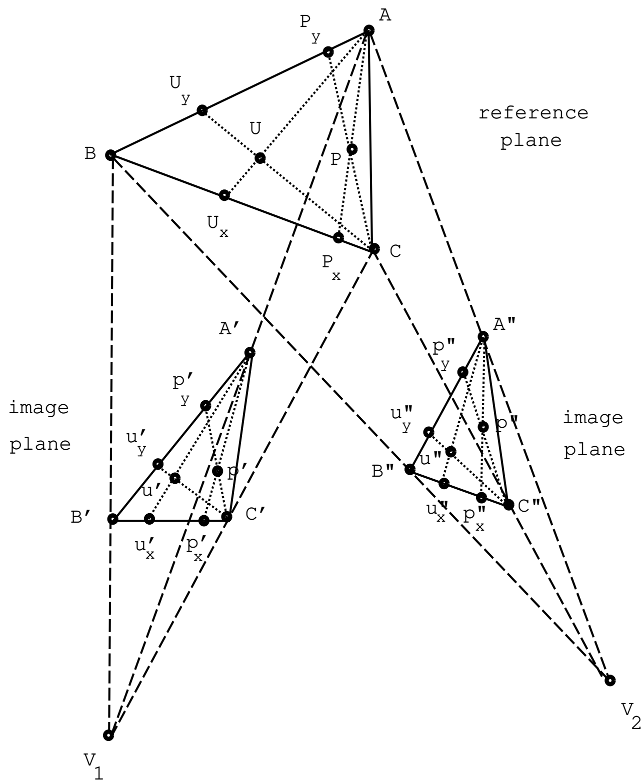


Figure A.1: The geometry underlying plane projectivity from four points.

equal to the cross ratio of B', U'_x, P'_x, C', i.e.

$$\frac{BC \cdot U_x P_x}{BP_x \cdot U_x C} = \frac{B'C' \cdot U'_x P'_x}{B'P'_x \cdot U'_x C'}.$$

This form of cross ratio is known as the *canonical cross ratio*. In general there are 24 cross ratios, six of which are numerically different (see Appendix B for more details on cross-ratios). Similarly, the cross ratio along the y-axis of the reference frame is equal to the cross ratio of the corresponding points in both views.

Therefore, for any point p' in the first view, we construct its x and y locations, p'_x, p'_y, along $B'C'$ and $B'A'$, respectively. From the equality of cross ratios we find the locations of p''_x, p''_y, and that leads to p''. Because we have used only projective constructions, i.e. straight lines project to straight lines, we are guaranteed that p' and p'' are corresponding points.

Algebraic Derivation

From an algebraic point of view it is convenient to view points as laying on rays emanating from the center of projection. A ray representation is also called the *homogeneous coordinates* representation of the plane, and is achieved by adding a third coordinate. Two vectors represent the same point $X = (x, y, z)$ if they differ at most by a scale factor (different locations along the same ray). A key result, which makes this representation amenable to application of linear algebra to geometry, is described in the following proposition:

Proposition 11 *A projectivity of the plane is equivalent to a linear transformation of the homogeneous representation.*

The proof is omitted here, and can be found in Tuller (1967, Theorems 5.22, 5.24). A projectivity is equivalent, therefore, to a linear transformation applied to the rays. Because the correspondence between points and coordinates is not one-to-one, we have to take scalar factors of proportionality into account when representing a projective transformation. An arbitrary projective transformation of the plane can be represented as a non-singular linear transformation (also called *collineation*) $\rho X' = TX$, where ρ is an arbitrary scale factor.

Given four corresponding rays $p_j = (x_j, y_j, 1) \longleftrightarrow p'_j = (x'_j, y'_j, 1)$, we would like to find a linear transformation T and the scalars ρ_j such that $\rho_j p'_j = Tp_j$. Note that because only ratios are involved, we can set $\rho_4 = 1$. The following are a basic lemma and theorem adapted from Semple and Kneebone (1952).

Lemma 1 *If $p_1, ..., p_4$ are four vectors in R^3, no three of which are linearly dependent, and if $e_1, ..., e_4$ are respectively the vectors $(1,0,0), (0,1,0), (0,0,1), (1,1,1)$, there exists a non-singular linear transformation A such that $Ae_j = \lambda_j p_j$, where the λ_j are non-zero scalars; and the matrices of any two transformations with this property differ at most by a scalar factor.*

Proof: Let p_j have the components $(x_j, y_j, 1)$, and without loss of generality let $\lambda_4 = 1$. The matrix A satisfies three conditions $Ae_j = \lambda_j p_j$, $j = 1, 2, 3$ if and only if $\lambda_j p_j$ is the j'th column of A. Because of the fourth condition, the values $\lambda_1, \lambda_2, \lambda_3$ satisfy

$$[p_1, p_2, p_3] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = p_4$$

and since, by hypothesis of linear independence of p_1, p_2, p_3, the matrix $[p_1, p_2, p_3]$ is non-singular, the λ_j are uniquely determined and non-zero. The matrix A is therefore determined up to a scalar factor. $\square$

Theorem 4 *If $p_1, ..., p_4$ and $p'_1, ..., p'_4$ are two sets of four vectors in R^3, no three vectors in either set being linearly dependent, there exists a non-singular linear transformation T such that $Tp_j = \rho_j p'_j$ $(j = 1, ..., 4)$, where the ρ_j are scalars; and the matrix T is uniquely determined apart from a scalar factor.*

Proof: By the lemma, we can solve for A and λ_j that satisfy $Ae_j = \lambda_j p_j$ $(j = 1, ..., 4)$, and similarly we can choose B and μ_j to satisfy $Be_j = \mu_j p'_j$; and without loss of generality assume that $\lambda_4 = \mu_4 = 1$. We then have, $T = BA^{-1}$ and $\rho_j = \frac{\mu_j}{\lambda_j}$. If, further, $Tp_j = \rho_j p'_j$ and $Up_j = \sigma_j p'_j$, then $TAe_j = \rho_j \lambda_j p'_j$ and $UAe_j = \sigma_j \lambda_j p'_j$; and therefore, by the lemma, $TA = \tau UA$, i.e., $T = \tau U$ for some scalar τ. $\square$

The immediate implication of the theorem is that one can solve directly for T and ρ_j $(\rho_4 = 1)$. Four points provide twelve equations and we have twelve unknowns (nine for T and three for ρ_j). Furthermore, because the system is linear, one can look for a least squares solution by using more than four corresponding points (they all have to be coplanar): each additional point provides three more equations and one more unknown (the ρ associated with it).

Alternatively, one can eliminate ρ_j from the equations, set $T_{3,3} = 1$ and set up directly a system of eight linear equations as follows. In general we have four corresponding rays $p_j = (x_j, y_j, z_j) \longleftrightarrow p'_j = (x'_j, y'_j, z'_j)$, $j = 1, ..., 4$, and the linear transformation T satisfies $\rho_j p'_j = Tp_j$. By eliminating ρ_j, each pair of corresponding rays contributes the following two linear equations:

$$x_j t_{1,1} + y_j t_{1,2} + z_j t_{1,3} - \frac{x_j x'_j}{z'_j} t_{3,1} - \frac{y_j x'_j}{z'_j} t_{3,2} = \frac{z_j x'_j}{z'_j}$$
$$x_j t_{2,1} + y_j t_{2,2} + z_j t_{2,3} - \frac{x_j y'_j}{z'_j} t_{3,1} - \frac{y_j y'_j}{z'_j} t_{3,2} = \frac{z_j y'_j}{z'_j}$$

A similar pair of equations can be derived in the case $z'_j = 0$ (ideal points) by using either x'_j or y'_j (all three cannot be zero).

Projectivity Between Two image Planes of an Uncalibrated Camera

We can use the fundamental theorem of plane projectivity to recover the projective transformation that was illustrated geometrically in Figure A.1. Given four corresponding points $(x_j, y_j) \longleftrightarrow (x'_j, y'_j)$ that are projected from four coplanar points in space we would like to find the projective transformation A that accounts for all other correspondences $(x, y) \longleftrightarrow (x', y')$ that are projected from coplanar points in space.

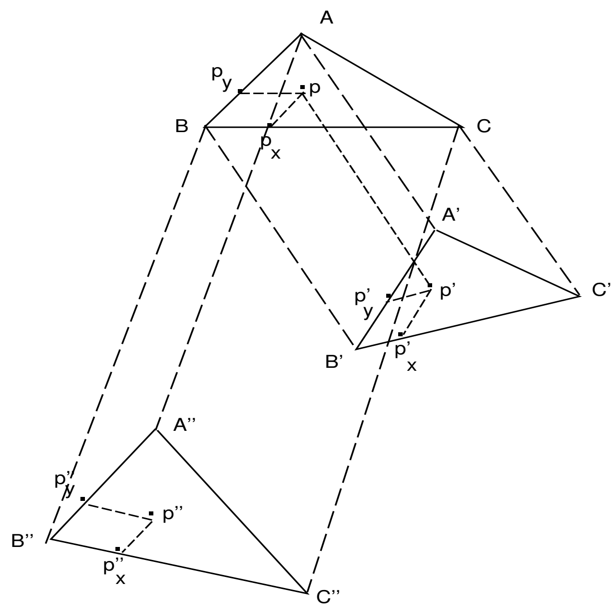


Figure A.2: Setting a projectivity under parallel projection.

The standard way to proceed is to assume that both image planes are parallel to their xy plane with a focal length of one unit, or in other words to embed the image coordinates in a 3D vector whose third component is 1. Let $p_j = (x_j, y_j, 1)$ and $p'_j = (x'_j, y'_j, 1)$ be the the chosen representation of image points. The true coordinates of those image points may be different (if the image plane are in different positions than assumed), but the main point is that all such representations are projectively equivalent to each other. Therefore, $\rho_j p_j = B\hat{p}_j$ and $\mu_j p'_j = C\hat{p}'_j$, where $\hat{p}_j$ and $\hat{p}'_j$ are the true image coordinates of these points. If T is the projective transformation determined by the four corresponding points $\hat{p}_j \longleftrightarrow \hat{p}'_j$, then $A = CTB^{-1}$ is the projective transformation between the assumed representations $p_j \longleftrightarrow p'_j$.

Therefore, the matrix A can be solved for directly from the correspondences $p_j \longleftrightarrow p'_j$ (the system of eight equations detailed in the previous section). For any given point $p = (x, y, 1)$, the corresponding point $p' = (x', y', 1)$ is determined by Ap followed by normalization to set the third component back to 1.

A.1 Plane Projectivity in Affine Geometry

In parallel projection we can take advantage of the fact that parallel lines project to parallel lines. This allows to define coordinates on the plane by subtending lines parallel to the axes (see Figure A.2). Note also that the two trapezoids $BB'p'_xp_x$ and $BB'C'C$ are similar trapezoids, therefore,

$$\frac{BC}{p_xC} = \frac{B'C'}{p'_xC'}.$$

This provides a geometric derivation of the result that three points are sufficient to set up a projectivity between any two planes under parallel projection.

Algebraically, a projectivity of the plane can be uniquely represented as a 2D affine transformation of the non-homogeneous coordinates of the points. Namely, if $p = (x, y)$ and $p' = (x', y')$ are two corresponding points, then

$$p' = Ap + w$$

where A is a non-singular matrix and w is a vector. The six parameters of the transformation can be recovered from two non-collinear sets of three points, p_o, p_1, p_2 and p'_o, p'_1, p'_2. Let

$$A = \begin{bmatrix} x'_1 - x'_o, x'_2 - x'_o \\ y'_1 - y'_o, y'_2 - y'_o \end{bmatrix} \begin{bmatrix} x_1 - x_o, x_2 - x_o \\ y_1 - y_o, y_2 - y_o \end{bmatrix}^{-1}$$

and $w = p'_o - Ap_o$, which together satisfy $p'_j - p'_o = A(p_j - p_o)$ for $j = 1, 2$. For any arbitrary point p on the plane, we have that p is spanned by the two vectors $p_1 - p_o$ and $p_2 - p_o$, i.e., $p - p_o = \alpha_1(p_1 - p_o) + \alpha_2(p_2 - p_o)$; and because translation in depth is lost in parallel projection, we have that $p' - p'_o = \alpha_1(p'_1 - p'_o) + \alpha_2(p'_2 - p'_o)$, and therefore $p' - p'_o = A(p - p_o)$.

Cross-Ratio and the Linear Combination of Rays

Appendix B

The cross-ratio of four collinear points A, B, C, D is preserved under central projection and is defined as:

$$\alpha = \frac{AB}{AC} \div \frac{DB}{DC} = \frac{A'B'}{A'C'} \div \frac{D'B'}{D'C'},$$

(see Figure B.1). All permutations of the four points are allowed, and in general there are six distinct cross-ratios that can be computed from four collinear points. Because the cross-ratio is invariant to projection, any transversal meeting four distinct concurrent rays in four distinct points will have the same cross ratio — therefore one can speak of the cross-ratio of rays (concurrent or parallel) $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$.

The cross-ratio result in terms of rays, rather than points, is appealing for the reasons that it enables the application of linear algebra (rays are represented as points in homogeneous coordinates), and more important, enables us to treat ideal points as any other point (critical for having an algebraic system that is well defined under both central and parallel projection).

The cross-ratio of rays is computed algebraically through linear combination of points in homogeneous coordinates (see Gans 1969, pp. 291–295), as follows. Let the the rays $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$ be represented by vectors $(a_1, a_2, a_3), ..., (d_1, d_2, d_3)$, respectively. We can represent the rays $\boldsymbol{a}, \boldsymbol{d}$ as a linear combination of the rays $\boldsymbol{b}, \boldsymbol{c}$, by

$$\boldsymbol{a} = \boldsymbol{b} + k\boldsymbol{c}$$

$$\boldsymbol{d} = \boldsymbol{b} + k'\boldsymbol{c}$$

For example, k can be found by solving the linear system of three equation $\rho\boldsymbol{a} = \boldsymbol{b} + k\boldsymbol{c}$ with two unknowns ρ, k (one can solve using any two of the three equations, or find a least squares solution using all three equations). We shall assume, first, that the points are Euclidean. The ratio in which A divides the line BC can be derived by:

$$\frac{AB}{AC} = \frac{\frac{a_1}{a_3} - \frac{b_1}{b_3}}{\frac{a_1}{a_3} - \frac{c_1}{c_3}} = \frac{\frac{b_1+kc_1}{b_3+kc_3} - \frac{b_1}{b_3}}{\frac{b_1+kc_1}{b_3+kc_3} - \frac{c_1}{c_3}} = -k\frac{c_3}{b_3}$$

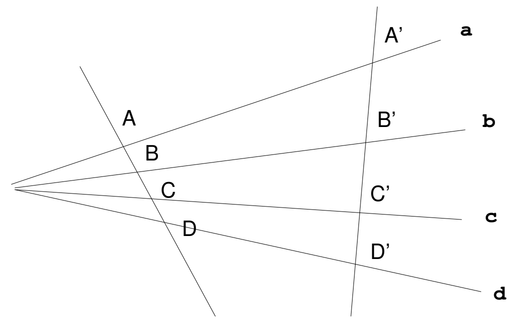


Figure B.1: The cross-ratio of four distinct concurrent rays is equal to the cross-ratio of the four distinct points that result from intersecting the rays by a transversal.

Similarly, we have $\frac{DB}{DC} = -k'\frac{c_3}{b_3}$ and, therefore, the cross-ratio of the four rays is $\alpha = \frac{k}{k'}$. The same result holds under more general conditions, i.e., points can be ideal as well:

Proposition 12 *If A, B, C, D are distinct collinear points, with homogeneous coordinates $\boldsymbol{b} + k\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{b} + k'\boldsymbol{c}$, then the canonical cross-ratio is $\frac{k}{k'}$.*

(for a complete proof, see Gans 1969, pp. 294–295). For our purposes it is sufficient to consider the case when one of the points, say the vector $\boldsymbol{d}$, is ideal (i.e. $d_3 = 0$). From the vector equation $\rho\boldsymbol{d} = \boldsymbol{b} + k'\boldsymbol{c}$, we have that $k' = -\frac{b_3}{c_3}$ and, therefore, the ratio $\frac{DB}{DC} = 1$. As a result, the cross-ratio is determined only by the first term, i.e., $\alpha = \frac{AB}{AC} = k$ — which is what we would expect if we represented points in the Euclidean plane and allowed the point D to extend to infinity along the line A, B, C, D (see Figure B.1).

The derivation so far can be translated directly to our purposes of computing the projective shape constant by replacing $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$ with $p', \tilde{p}', \hat{p}', V_l$, respectively.

On Epipolar Transformations

Appendix C

Proposition 13 *The epipolar lines pV_r and $p'V_l$ are perspectively related.*

Proof: Consider Figure C.1. We have already established that p projects onto the left epipolar line $p'V_l$. By definition, the right epipole V_r projects onto the left epipole V_l, therefore, because lines are projective invariants the line pV_r projects onto the line $p'V_l$. $\square$

The result that epipolar lines in one image are perspectively related to the epipolar lines in the other image, implies that there exists a projective transformation F that maps epipolar lines l_j onto epipolar lines l'_j, that is $Fl_j = \rho_j l'_j$, where $l_j = p_j \times V_r$ and $l'_j = p'_j \times V_l$. From the property of point/line duality of projective geometry (Semple and Kneebone, 1952), the transformation E that maps points on left epipolar lines onto points on the corresponding right epipolar lines is induced from F, i.e., $E = (F^{-1})^t$.

Proposition 14 (point/line duality) *The transformation for projecting p onto the left epipolar line $p'V_l$, is $E = (F^{-1})^t$.*

Proof: Let l, l' be corresponding epipolar lines, related by the equation $\rho l' = Fl$. Let p, p' be any two points, one on each epipolar line (not necessarily corresponding points). From the point/line incidence axiom we have that $l^t \cdot p = 0$. By substituting l we have

$$\left[\rho F^{-1} l'\right]^t \cdot p = 0 \qquad \Longrightarrow \qquad \rho l'^t \cdot \left[F^{-t} p\right] = 0.$$

Therefore, the collineation $E = (F^{-1})^t$ maps points p onto the corresponding left epipolar line. $\square$

It is intuitively clear that the epipolar line transfomation F is not unique, and therefore the induced transformation E is not unique either. The correspondence between the epipolar lines is not disturbed under translation along the line V_1V_2, or under non-rigid camera motion that results from tilting the image plane with respect to the optical axis such that the epipole remains on the line V_1V_2.

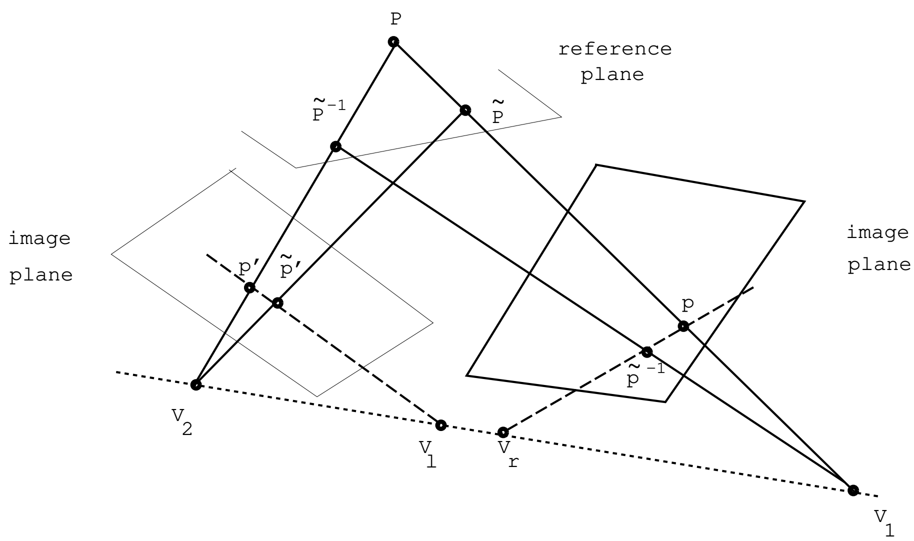


Figure C.1: Epipolar lines are perspectively related.

Proposition 15 *The epipolar transformation F is not unique.*

Proof: A projective transformation is determined by four corresponding pencils. The transformation is unique (up to a scale factor) if no three of the pencils are linearly dependent, i.e., if the pencils are lines, then no three of the four lines should be coplanar. The epipolar line transformation F can be determined by the corresponding epipoles, $V_r \longleftrightarrow V_l$, and three corresponding epipolar lines $l_j \longleftrightarrow l'_j$. We show next that the epipolar lines are coplanar, and therefore, F cannot be determined uniquely.

Let p_j and p'_j, $j = 1, 2, 3$, be three corresponding points and let $l_j = p_j \times V_r$ and $l'_j = p'_j \times V_l$. Let $\bar{p}_3 = \alpha p_1 + \beta p_2$, $\alpha + \beta = 1$, be a point on the epipolar line $p_3 V_r$ collinear with p_1, p_2. We have,

$$l_3 = p_3 \times V_r = (a\bar{p}_3 + bV_r) \times V_r = a\bar{p}_3 \times V_r = a\alpha l_1 + a\beta l_2,$$

and similarly $l'_3 = \alpha' l'_1 + \beta' l'_2$. $\square$

Computational Background on Image Formation

Appendix D

D.1 The Standard Model of Image Formation

Image formation in a biological system is formed by the response of retinal photo-receptors, called *cones*, to incoming light from the scene, also referred to as *scene radiance*. Retinal cones come in three types which vary in their spectral sensitivity, or how the absorption of light varies with wavelength. The peak sensitivity of the three types of cones in the human retina lie in the violet, the green and the yellow-green, respectively (also referred to as short-wave, middle-wave and long-wave receptors). Similarly, when an image is captured with a CCD camera, filters of different spectral sensitivity are used, often Red, Green and Blue filters, to form the image (which is composed of three images, one per filter). The quantum catch or the measured signal from a cone or CCD filter is given by

$$I_k(p) = \int_{\lambda_1}^{\lambda_2} S(\lambda)\rho_p(\lambda)G(\boldsymbol{s}, \boldsymbol{v}, \boldsymbol{n}_p)R_k(\lambda)d\lambda,$$

where $k = 1, 2, 3$ represents the cone type or the filter type, and $I_k(p)$ is the image signal (*image irradiance*) associated with filter k at image location p. The wavelengths $\lambda_1 = 400nm$ and $\lambda_2 = 700nm$ cover the range of the visible spectrum.

The function $R_k(\lambda)$ represents the spectral sensitivity of the k'th cone or filter, and it is a function of wavelength alone, there is no dependence on spatial location. The product $L(\lambda, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{n}_p) = S(\lambda)\rho_p(\lambda)G(\boldsymbol{s}, \boldsymbol{v}, \boldsymbol{n}_p)$ represents the scene radiance and it is a function of illumination S, surface reflectance ρ and the viewing and scene geometry G. The illumination is composed of light sources, that have a direction in space, represented by vector $\boldsymbol{s}$, and a spectral power distribution $S(\lambda)$, i.e. the intensity of light as a function of wavelength. For simplicity, light sources are assumed to be located relatively far away from the scene, therefore the light rays arriving from each source meet the surface in parallel rays (*point* light sources). The surface reflectance function $\rho_p(\lambda)$ represents the percentage of light reflected back as a function of wavelength at position p. Surface reflectance is also called *albedo* and depends on surface material and surface color. The geometric component G depends on the direction of light sources $\boldsymbol{s}$, the viewer's direction $\boldsymbol{v}$, and the surface normal $\boldsymbol{n}_p$ at the scene point P that is projecting to image point p (model of projection is not important here). An important assumption with the standard model is that the effect of multiple light sources is additive, and therefore it is mathematically convenient to assume a single point light source when writing down the image formation equation. The detailed relationship $S(\lambda)\rho_p(\lambda)G(\boldsymbol{s}, \boldsymbol{v}, \boldsymbol{n}_p)$ is referred to as the *reflectance model* and is described below.

D.2 The Standard Reflectance Model

The standard reflectance model applies to inhomogeneous 'rough' surfaces, such as plastic, paint and many dielectric surfaces. These reflectance models rely on the application of geometric optics which holds under the assumption that the wavelength of light is much smaller than the *roughness* of the surface, or to the dimensions of the microscopic surface undulations. Geometric optics models such as the Torrance-Sparrow (1967), or the Trowbridge-Reitz (1975) provide a good approximation for shiny smooth materials that would otherwise require the application of physical optics, based on the electromagnetic wave theory, to provide an exact model (Beckmann and Spizzichino, 1963).

An optically inhomogeneous material consists of carrier material, which is largely transparent, and of pigment particles embedded in the carrier. The light that is reflected from such a surface is, in general, composed of two types: a diffuse component, referred to as the *Lambertian* component or the *body reflection*, and a *specular* or *interface* component (Shafer, 1985). When light reaches the surface some portion of it is refracted into the carrier where it is scattered from the pigment particles. Some of the scattered rays find their way back to the surface in a variety of directions, resulting in diffuse reflection. Depending on the pigment material and its distribution, the diffuse component undergoes a spectral change which is represented by the product of the spectral composition function of the light source and the albedo of the surface. Therefore, the diffuse component carries the color of the surface (together with the color of the illuminant). Another property of the diffuse component is the Lambertian property due to the randomness of the re-emitted light that is scattered by the pigments. The Lambertian property is that the amount of reflected light does not depend on the viewing direction, but only on the cosine angle between the incidence light ray and the normal to the surface, i.e.

$$L(\lambda, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{n}_p) = L(\lambda, \boldsymbol{s}, \boldsymbol{n}_p) = S(\lambda)\rho_p(\lambda)\boldsymbol{n}_p \cdot \boldsymbol{s},$$

where $\boldsymbol{n}_p \cdot \boldsymbol{s}$ is the dot product between the unit vectors $\boldsymbol{n}_p$ and $\boldsymbol{s}$, representing the normal direction at surface point P and the direction of the light source, respectively. The image irradiance due to the Lambertian component becomes

$$I_k(p) = \left[\int_{\lambda_1}^{\lambda_2} S(\lambda)\rho_p(\lambda)R_k(\lambda)d\lambda\right]\boldsymbol{n}_p \cdot \boldsymbol{s} = \mu_k(p)\boldsymbol{n}_p \cdot \boldsymbol{s}, \tag{D.1}$$

where $k = 1, 2, 3$ represents the CCD filter, i.e. the R,G,B filters. The second component of the surface reflectance is due to external scatter, or reflection from the air-surface interface, and is narrowly diffused around a single direction, called the *specular direction*. The external scattering depend on the roughness of the surface; light rays are reflected between the surface's micro-facets before they are scattered into space. The smoother the surface the less scattering occurs and the more reflections in the specular direction (making an equal angle of incidence around the surface normal). The roughness of the surface determines, therefore, the scatter around the specular direction, which is also referred to as the *specular lobe*, or the *forescatter lobe*. For a perfectly smooth surface, like a mirror, there is no scattering and the specular lobe turns into a *specular spike*. A simple model of the specular lobe, using geometric optics, is the microscopic facet model which goes as follows: A rough surface is modeled as being made up of microscopically planar reflectors that are inclined randomly about the mean surface. The distribution of facets about the mean causes the reflected flux to distribute around the specular direction. Accurate mathematical descriptions of the shape of the specular lobe can be made from such a facet model (Torrance and Sparrow 1967, Phong 1975, Trowbridge and Reitz 1975). The specular component due to Phong's model has the form

$$F(\lambda, \boldsymbol{n}_p, \boldsymbol{v}, \boldsymbol{s}) = \beta S(\lambda)(\boldsymbol{n}_p \cdot \boldsymbol{h})^c,$$

where $\boldsymbol{h}$ is a bi-sector of the vectors pointing to the viewer and to the light source, $c \approx 50$ is a constant that represents the degree of sharpness or extent of scatter around the specular direction, and β is a fixed constant. Note that the color of the specular reflection is the same is the color of the light source and this is because the index of refraction of the carrier is constant with respect to wavelength and is independent of the imaging geometry (this is generally not true for homogeneous surfaces). The overall image irradiance is a linear combination of Lambertian component and the specular component, and has the form

$$I_k(p) = \mu_k(p)\boldsymbol{n}_p \cdot \boldsymbol{s} + \nu_k(\boldsymbol{n}_p \cdot \boldsymbol{h})^c. \tag{D.2}$$

The specular term takes different forms depending on the reflectance model, and the one we used here (Phong's model) is the simplest. The main point is that for rough surfaces, such as paint, paper, plastic and so forth, the reflectance is dominantly Lambertian because the specular reflection falls off exponentially from the specular direction. Therefore, if the surface is not flat we expect the specular reflections to occupy only small regions in the image, and the rest is dominated by diffuse reflection. The approach we take is to assume Lambertian reflection as the model of surface reflection and deal with the specularities separately, by detecting and removing them from the image.

Bibliography

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2:284–299, 1985.

[2] G. Adiv. Inherent ambiguities in recovering 3D motion and structure from a noisy flow field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(5):477–489, 1989.

[3] J. Aloimonos and C.M. Brown. On the kinetic depth effect. *Biological Cybernetics*, 60:445–455, 1989.

[4] P. Anandan. A unified perspective on computational techniques for the measurement of visual motion. In *Proceedings Image Understanding Workshop*, pages 219–230, Los Angeles, CA, February 1987. Morgan Kaufmann, San Mateo, CA.

[5] I.A. Bachelder and S. Ullman. Contour matching using local affine transformations. In *Proceedings Image Understanding Workshop*. Morgan Kaufmann, San Mateo, CA, 1992.

[6] H.S. Baird. *Model-Based image matching using location.* MIT Press, Cambridge, MA, 1985.

[7] D.H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13:111–122, 1981.

[8] J. Barron, D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt. Performance of optical flow techniques. Technical Report 299, University of Western Ontario, October 1991.

[9] P. Beckmann and A. Spizzichino. *The scattering of electromagnetic waves from rough surfaces.* Pergamon, New-York, 1963.

[10] J.R. Bergen and E.H. Adelson. Hierarchical, computationally efficient motion estimation algorithm. *Journal of the Optical Society of America*, 4:35, 1987.

[11] J.R. Bergen and R. Hingorani. Hierarchical motion-based frame rate conversion. Technical report, David Sarnoff Research Center, 1990.

[12] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.

[13] T.O. Binford. Visual perception by computer. In *Proc. IEEE Conf. on Systems and Control*, Miami, FL, 1971.

[14] T.O. Binford. Survey of model-based image analysis systems. *International Journal of Robotics Research*, 1:18–64, 1982.

[15] R.C. Bolles and R.A. Cain. Recognizing and locating partially visible objects: the local feature focus method. *International Journal of Robotics Research*, 1(3):57–82, 1982.

[16] T.M. Breuel. *Geometric apects of visual object recognition*. PhD thesis, M.I.T Artificial Intelligence Laboratory, May 1992.

[17] T. Broida, S. Chandrashekhar, and R. Chellapa. recursive 3-d motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, 26:639–656, 1990.

[18] M.J. Brooks and B.K.P. Horn. Shape and source from shading. In *Proceedings IJCAI*, pages 932–936, Los Angeles, CA, 1985.

[19] R. Brooks. Symbolic reasoning among 3-dimensioanl models and 2-dimensional images. *Artificial Intelligence*, 17:285–349, 1981.

[20] D.C. Brown. Close-range camera calibration. *Photogrammetric Engineering*, 37:855–866, 1971.

[21] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communication*, 31:532–540, 1983.

[22] J. Canny. Finding edges and lines in images. A.I. TR No. 720, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1983.

[23] T.A. Cass. A robust implementation of 2D model-based recognition. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, Ann Arbor, Michigan, 1988.

[24] T.A. Cass. Polynomial-time object recognition in the presence of clutter, occlusion, and uncertainty. In *Proceedings of the European Conference on Computer Vision*, pages 834–842, Santa Margherita Ligure, Italy, June 1992.

[25] P. Cavanagh. What's up in top-down processing? In *Proceedings of the XIIIth ECVP, Andrei Gorea (Ed.)*, pages 1–10, 1990.

[26] J.H. Connell. Learning shape descriptions: generating and generalizing models of visual objects. A.I. TR No. 853, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1985.

[27] J.E. Cutting. *Perception with an eye for motion.* MIT Press, Cambridge, MA, 1986.

[28] L.S. Davis. Shape matching using relaxation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1:60–72, 1979.

[29] R.J. Douglass. Interpreting three dimensional scenes: A model building approach. *Computer Graphics and Image Processing*, 17:91–113, 1981.

[30] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis.* John Wiley, New York, 1973.

[31] R. Dutta and M.A. Synder. Robustness of correspondence based structure from motion. In *Proceedings of the International Conference on Computer Vision*, pages 106–110, Osaka, Japan, December 1990.

[32] S. Edelman and H.H. Bulthoff. Viewpoint-specific representations in three-dimensional object recognition. A.I. Memo No. 1239, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[33] Jr. E.N. Coleman and R. Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics and Image Processing*, 18:309–328, 1982.

[34] W. Faig. Calibration of close-range photogrammetry systems: Mathematical formulation. *Photogrammetric Engineering and Remote Sensing*, 41:1479–1486, 1975.

[35] M.J. Farah. *Visual Agnosia.* MIT Press, Cambridge, MA, 1990.

[36] O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proceedings of the European Conference on Computer Vision*, pages 563–578, Santa Margherita Ligure, Italy, June 1992.

[37] O.D. Faugeras and M. Hebert. The representation, recognition and location of 3D objects. *International Journal of Robotics Research*, 5(3):27–52, 1986.

[38] O.D. Faugeras, Q.T. Luong, and S.J. Maybank. Camera self calibration: Theory and experiments. In *Proceedings of the European Conference on Computer Vision*, pages 321–334, Santa Margherita Ligure, Italy, June 1992.

[39] O.D. Faugeras and S. Maybank. Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, 4:225–246, 1990.

[40] M.A. Fischler and R.C. Bolles. Random sample concensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, 1981.

[41] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13, 1991.

[42] D. Gans. *Transformations and Geometries.* Appleton-Century-Crofts, New York, 1969.

[43] A.L. Gilchrist. The perception of surface blacks and whites. *SIAM J. Comp.*, pages 112–124, 1979.

[44] F. Glazer, G. Reynolds, and P. Anandan. Scene matching through hierarchical correlation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 432–441, Washington, D.C., 1983.

[45] W. E. L. Grimson. A computational theory of visual surface interpolation. *Proceedings of the Royal Society of London B*, 298:395–427, 1982.

[46] W.E.L. Grimson and D.P. Huttenlocher. On the sensitivity of Hough transform for object recognition. In *Proceedings of the International Conference on Computer Vision*, pages 700–706, Tampa, FL, Dec. 1988.

[47] W.E.L. Grimson, D.P. Huttenlocher, and D.W. Jacobs. A study of affine matching-with bounded sensor error. In *Proceedings of the European Conference on Computer Vision*, pages 291–306, Santa Margherita Ligure, Italy, June 1992. Also in M.I.T AI Memo No. 1250, Aug. 1991.

[48] W.E.L. Grimson and T. Lozano-Pérez. Model-based recognition and localization from sparse data. *International Journal of Robotics Research*, 3:3–35, 1984.

[49] R. Hartley, R. Gupta, and Tom Chang. Stereo from uncalibrated cameras. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 761–764, Champaign, IL., June 1992.

[50] D.J. Heeger. Optical flow from spatiotemporal filters. In *Proceedings Image Understanding Workshop*, pages 181–190, Los Angeles, CA, February 1987. Morgan Kaufmann, San Mateo, CA.

[51] E.C. Hildreth. Computations underlying the measurement of visual motion. *Artificial Intelligence*, 23(3):309–354, August 1984.

[52] E.C. Hildreth. Recovering heading for visually-guided navigation. A.I. Memo No. 1297, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, June 1991.

[53] E. Hille. *Analytic function theory.* Ginn/Blaisdell, Waltham, MA, 1962.

[54] D. Hoffman and W. Richards. Parts of recognition. In A.P. Pentland, editor, *From pixels to predicates.* Ablex, Norwood, NJ, 1986.

[55] B.K.P. Horn. Image intensity understanding. *Artificial Intelligence*, 8:201–231, 1977.

[56] B.K.P. Horn. *Robot Vision.* MIT Press, Cambridge, Mass., 1986.

[57] B.K.P. Horn. Relative orientation. *International Journal of Computer Vision*, 4:59–78, 1990.

[58] B.K.P. Horn. Relative orientation revisited. *Journal of the Optical Society of America*, 8:1630–1638, 1991.

[59] B.K.P. Horn and M.J. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33:174–208, 1986.

[60] B.K.P. Horn and B.G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[61] P. Van Hove. Model based silhouette recognition. In *Proceedings of the IEEE Workshop on Computer Vision*, 1987.

[62] T.S. Huang and C.H. Lee. Motion and structure from orthographic projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:536–540, 1989.

[63] D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Proceedings of the Royal Society of London*, 160:106–154, 1962.

[64] D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proceedings of the International Conference on Computer Vision*, pages 102–111, London, December 1987.

[65] K. Ikeuchi and B.K.P. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17:141–184, 1981.

[66] D.W. Jacobs. Space efficient 3D model indexing. In *Proceedings Image Understanding Workshop*. Morgan Kaufmann, San Mateo, CA, January 1992. Also in M.I.T. AI memo 1353.

[67] P. Jolicoeur. The time to name disoriented natural objects. *Memory and Cognition*, 13:289–303, 1985.

[68] B. Julesz. *Foundations of Cyclopean Perception*. University of Chicago Press, Chicago, IL, 1971.

[69] M. Kinsbourne and E.K. Warrington. A disorder of simultaneous form perception. *Brain*, 85:461–486, 1962.

[70] G.J. Klinker, S.A. Shafer, and T. Kanade. A physical approach to color image understanding. *International Journal of Computer Vision*, 4:7–38, 1990.

[71] D.C. Knill and D. Kersten. Apparent surface curvature affects lightness perception. *Nature*, 351:228–230, 1991.

[72] J.J. Koenderink and A.J. Van Doorn. Photometric invariants related to solid shape. *Optica Acta*, 27:981–986, 1980.

[73] J.J. Koenderink and A.J. Van Doorn. Affine structure from motion. *Journal of the Optical Society of America*, 8:377–385, 1991.

[74] Y. Lamdan, J.T. Schwartz, and H.J. Wolfson. Object recognition by affine invariant matching. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 335–344, Ann Arbor, Michigan, 1988.

[75] E.H. Land and J.J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971.

[76] C.H. Lee. Structure and motion from two perspective views via planar patch. In *Proceedings of the International Conference on Computer Vision*, pages 158–164, Tampa, FL, December 1988.

[77] R.K. Lenz and R.Y. Tsai. Techniques for calibration of the scale factor and image center for high accuracy 3D machine vision metrology. In *fourth int. Symp. on Robotics Research*, pages 68–75, Santa Cruz, CA, Aug 1987.

[78] B. Logan. Information in the zero-crossings of band pass signals. *Bell Syst. Tech. J.*, 56:510, 1977.

[79] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[80] H.C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proceedings of the Royal Society of London B*, 208:385–397, 1980.

[81] D.G. Lowe. *Perceptual organization and visual recognition.* Kluwer Academic Publishing, Hingham, MA, 1985.

[82] D.G. Lowe. Three-dimensional object recognition from single two dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[83] B.D. Lucas. *Generalized image matching by the method of differences.* PhD thesis, Dept. of Computer Science, Carnegie-Mellon University, 1984.

[84] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings IJCAI*, pages 674–679, Vancouver, 1981.

[85] R.K. Luneburg. *Mathematical Analysis of Binocular Vision.* Princeton University Press, Princeton, NJ, 1947.

[86] J.V. Mahoney. Image chunking: defining spatial building blocks for scene analysis. Master's thesis, Department of EECS, M.I.T, AI-TR 980, 1986.

[87] L.T. Maloney and B. Wandell. A computational model of color constancy. *Journal of the Optical Society of America*, 1:29–33, 1986.

[88] D. Marr. Early processing of visual information. *Proceedings of the Royal Society of London B*, 175:483–534, 1976.

[89] D. Marr. *Vision.* W.H. Freeman and Company, San Francisco, CA, 1982.

[90] D. Marr and E.C. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London B*, 207:187–217, 1980.

[91] D. Marr and H.K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B*, 200:269–291, 1978.

[92] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204:301–328, 1979.

[93] G.J. Mitchison and S.P. McKee. interpolation in stereoscopic matching. *Nature*, 315:402–404, 1985.

[94] R. Mohr, L. Quan, F. Veillon, and B. Boufama. Relative 3D reconstruction using multiple uncalibrated images. Technical Report RT 84-IMAG, LIFIA — IRIMAG, France, June 1992.

[95] C.M. Mooney. Recognition of ambiguous and unambiguous visual configurations with short and longer expsures. *Brit. J. Psychol.*, 51:119–125, 1960.

[96] M.C. Morrone and D.C. Burr. Feature detection in human vision: a phase-dependant energy model. *Proceedings of the Royal Society of London B*, 235:221–245, 1988.

[97] Y. Moses. Face recognition. Interim Ph.D. thesis report, submitted to Dept. Applied Mathematics, Weizmann Institute of Science, Israel, August 1989.

[98] H.H. Nagel. On the estimation of optical flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987.

[99] K. Nakayama, S. Shimojo, and G.H. Silverman. Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception*, 18:55–68, 1989.

[100] S.E. Palmer. Structural aspects of visual similarity. *Memory and Cognition*, 6:91–97, 1978.

[101] D.E. Pearson, E. Hanna, and K. Martinez. Computer-generated cartoons. In H. Barlow, C. Blakemore, and M. Weston-Smith, editors, *Images and Understanding*. Cambridge University Press, New York, NY, 1990. Collection based on Rank Prize Funds' International Symposium, Royal Society 1986.

[102] P. Perona and J. Malik. Detecting an localizing edges composed of steps, peaks and roofs. In *Proceedings of the International Conference on Computer Vision*, Osaka, Japan, December 1990.

[103] D.I. Perret, P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.S. Head, A.D. Milner, and M.A. Jeeves. Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London B*, 223:293–317, 1985.

[104] B.T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18:311–317, 1975.

[105] S. Pinker. Visual cognition: An introduction. *Cognition*, 18:1–63, 1984.

[106] T. Poggio. 3D object recognition: on a result of Basri and Ullman. Technical Report IRST 9005-03, May 1990.

[107] K. Prazdny. 'capture' of stereopsis by illusory contours. *Nature*, 324:393, 1986.

[108] V.S. Ramachandran. Capture of stereopsis and apparent motion by illusory contours. *Perception and Psychophysics*, 39:361–373, 1986.

[109] V.S. Ramachandran and P. Cavanagh. subjective contours capture stereopsis. *Nature*, 317:527–530, 1985.

[110] V.S. Ramachandran and V. Inada. spatial phase and frequency in motion capture of random-dot patterns. *Spatial Vision*, 1:57–67, 1985.

[111] J.H. Rieger and D.T. Lawton. Processing differential image motion. *Journal of the Optical Society of America*, 2:354–360, 1985.

[112] J.W. Roach and J.K. Aggarwal. Computer tracking of objects moving in space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:127–135, 1979.

[113] L.G. Roberts. Machine perception of three-dimensional solids. In et. al. Tippett, editor, *Optical and electro-optical Information processing*. MIT Press, Cambridge, MA, 1965.

[114] I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293, 1987.

[115] I. Rock, J. DiVita, and R. Barbeito. The effect on form perception of change of orientation in the third dimension. *Journal of Experimental Psychology*, 7:719–733, 1981.

[116] F. Rosenblatt. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books, Washington, D.C., 1962.

[117] E.B. Saff and A.D. Snider. *Fundamentals of complex analysis*. Prentice-Hall, New-Jersey, 1976.

[118] J.P.H. Van Santen and G. Sperling. Elaborated Reichardt detectors. *Journal of the Optical Society of America*, 2:300–321, 1985.

[119] J. Sanz and T. Huang. Image representation by sign information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:729–738, 1989.

[120] H. Von Schelling. Concept of distance in affine geometry and its application in theories of vision. *Journal of the Optical Society of America*, 46:309–315, 1956.

[121] J.G. Semple and G.T. Kneebone. *Algebraic Projective Geometry*. Clarendon Press, Oxford, 1952.

[122] S.A. Shafer. Using color to separate reflection components. *COLOR research and applications*, 10:210–218, 1985.

[123] A. Shashua. Correspondence and affine shape from two orthographic views: Motion and Recognition. A.I. Memo No. 1327, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, December 1991.

[124] A. Shashua. Illumination and view position in 3D visual recognition. In S.J. Hanson J.E. Moody and R.P. Lippmann, editors, *Advances in Neural Information Processing*

Systems 4, pages 404–411. San Mateo, CA: Morgan Kaufmann Publishers, 1992. Proceedings of the fourth annual conference NIPS, Dec. 1991, Denver, CO.

[125] A. Shashua. Projective structure from two uncalibrated images: structure from motion and recognition. A.I. Memo No. 1363, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, September 1992.

[126] A. Shashua and S. Ullman. Structural saliency: the detection of globally salient structures using a locally connected network. In *Proceedings of the International Conference on Computer Vision*, pages 321–327, Tampa, FL, December 1988. Also in MIT AI-memo 1061.

[127] A. Shashua and S. Ullman. Grouping contours by iterated pairing network. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 335–341. San Mateo, CA: Morgan Kaufmann Publishers, 1991. Proceedings of the third annual conference NIPS, Dec. 1990, Denver, CO.

[128] R.N. Shepard and J. Metzler. Mental rotation: effects of dimensionality of objects and type of task. *Journal of Experimental Psychology: Human Perception and Performance*, 14:3–11, 1988.

[129] D. Shoham and S. Ullman. Aligning a model to an image using minimal information. In *Proceedings of the International Conference on Computer Vision*, pages 259–263, Tampa, FL, December 1988.

[130] P. Sinha. The perception of shading and reflectance. Master's thesis, Department of EECS, M.I.T, April 1992.

[131] J.B. Subirana-Vilanova and W. Richards. Perceptual organization, figure-ground, attention and saliency. A.I. Memo No. 1218, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, August 1991.

[132] M. Tarr and S. PInker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

[133] D. Thompson and J.L. Mundy. Three dimensional model matching from an unconstrained viewpoint. In *Proceedings of IEEE Conference on Robotics and Automation*, page 280, 1987.

[134] W. B. Thompson and S. T. Barnard. Low-level estimation and interpretation of visual motion. *IEEE Computer*, 14, August 1981.

[135] J.T. Todd and P. Bressan. The perception of 3D affine structure from minimal apparent motion sequences. *Perception and Psychophysics*, 48:419–430, 1990.

[136] C. Tomasi. *shape and motion from image streams: a factorization method.* PhD thesis, School of Computer Science, Carnagie Mellon University, Pittsburgh, PA, 1991.

[137] C. Tomasi and T. Kanade. Factoring image sequences into shape and motion. In *IEEE Workshop on Visual Motion*, pages 21–29, Princeton, NJ, September 1991.

[138] K. Torrance and E. Sparrow. Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America*, 57:1105–1114, 1967.

[139] T.S. Trowbridge and K.P. Reitz. Average irregularity representation of a roughened surface for ray reflection. *Journal of the Optical Society of America*, 65:531–536, 1975.

[140] R.Y. Tsai and T.S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surface. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:13–26, 1984.

[141] A. Tuller. *A modern introduction to geometries.* D. Van Nostrand Co., Princeton, NJ, 1967.

[142] S. Ullman. *The Interpretation of Visual Motion.* MIT Press, Cambridge and London, 1979.

[143] S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254, 1989. Also: in MIT AI Memo 931, Dec. 1986.

[144] S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13:992—1006, 1991. Also in M.I.T AI Memo 1052, 1989.

[145] A. Verri and T. Poggio. Motion field and optical flow: Qualitative properties. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:490–498, 1989.

[146] A.M. Waxman and S. Ullman. Surface structure and 3-D motion from image flow: a kinematic analysis. *International Journal of Robotics Research*, 4:72–94, 1985.

[147] A.M. Waxman and K. Wohn. Contour evolution, neighborhood deformation and global image flow: Planar surfaces in motion. *International Journal of Robotics Research*, 4(3):95–108, Fall 1985.

[148] D. Weinshall. Model based invariants for linear model acquisition and recognition. Technical Report RC 17705, IBM T.J. Watson Research Center, 1992.

[149] W.M. Wells. *Statistical Object Recognition.* PhD thesis, M.I.T Artificial Intelligence Laboratory, Novemebr 1992.

[150] R.J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19:139–144, 1980.

[151] R.J. Woodham, Y. Iwahori, and R.A. Barman. Photometric stereo: Lambertain reflectance and light sources with unknown direction and strength. Technical Report University of British Columbia, CS, TR-91-18, 1991.

[152] A.L. Yuille, D.S. Cohen, and P.W. Hallinan. Feature extraction from faces using deformable templates. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 104–109, San Diego, CA, 1989.