Class Based Recognition and Synthesis under Varying Illumination Conditions

Thesis for the M.Sc. Degree

by

Tammy Riklin-Raviv

Under the Supervision of
Dr. Amnon Shashua

Submitted to the Senate of the Hebrew University of Jerusalem (1999)

Acknowledgments

First and foremost, I'd like to thank my advisor *Dr. Amnon Shashua* for his guidance, direction and support. Amnon gave me the opportunity to benefit from his wide knowladge and experience, helped me to acquire tools to proceed my research, and to expose the work to the computer vision community. His ideas and fashion of thinking contributed considerably to this thesis.

Lots of thanks to *Prof. Shmuel Peleg* for directing the Computer Vision Lab. activities, while creating fertile and creative work environment, to *Prof. Michael Werman* for reading and providing comments on previous drafts of the work and to *Dr. Daphna Weinshall* for guiding me in my first steps in the lab.

I would also like to thank to my fellow members from the computer vision lab. especially to *Yoram Gdalyahu* for being a good friend, who encouraged and supported me all the way, to *Shai Avidan* for letting me enjoy his experience, endless patience and smart advises. Special thanks also to *Erez Shilat, Adi Shavit, Moshe Ben-Ezra, Yoni Wexler* and all the others for their help, encouragement and friendship.

Finally, I'd like to thank my husband, *Alon* for his love, help and encouragement which accompany me throughout the research.

Contents

Abstract

The thesis addresses the problem of "class-based" image-based recognition and rendering with varying illumination. The rendering problem is defined as follows: given a single input image of an object, and a sample of images with varying illumination conditions of other objects of the same general class, re-render the input image to simulate new illumination conditions. The class-based recognition problem is similarly defined: given a single image of an object in a database of images of other objects, some of them are multiply sampled under varying illumination, identify (match) any novel image of that object under varying illumination with the single image of that object in the database.

We focus on Lambertian surface classes, and in particular the class of human faces. The key result in our approach is based on a definition of an illumination invariant signature image which enables an analytic generation of the image space with varying illumination. We show that a small database of objects — in our experiments as few as two objects — is sufficient for generating the image space with varying illumination of any new object of the class from a single input image of that object. In many cases the recognition results outperform by far conventional methods and the re-rendering is of remarkable quality considering the size of the database of example images and the mild pre-process required for making the algorithm work.

Chapter 1

Introduction

Consider the image space generated by applying a source of variability, say changing illumination or changing viewing positions, on a $3D$ object or scene. Under certain circumstances the images generated by varying the parameters of the source can be represented as a function of a small number of sample images from the image space. For example, the image space of a $3D$ Lambertian surface is determined by a basis of three images, ignoring cast-shadows [Shashua, 1992, Shashua, 1997, Hallinan, 1994, Belhumeur et al., 1996, Nimeroff et al., 1994, Schoeneman et al., 1993]. In this case, the low dimensionality of the image space under lighting variations is useful for synthesizing novel images given a small number of model images, or in other words, provides the means for an "image-based rendering" process in which sampled images replace geometric entities formed by textured micro-polygons for rendering new images. Visual recognition and image re-rendering (synthesis) are intimately related. Recognizing a familiar object from a single picture under some source of variation requires a handle on how to capture the image space created by that source of variation. In other words, the process of visual recognition entails an ability to capture an equivalence class relationship that is either "generative", i.e., create a new image from a number of example images of an object, or "invariant", i.e., create a "signature" of the object that remains invariant under the source of variation under consideration. For example, in a generative process a set of basis images may form a compact representation of the image space. A novel input image is then considered part of the image space if it can be synthesized from the set of basis images. In a process based on invariance, on the other hand, the signature may be a "neutral" image, say the object under a canonical lighting condition or viewing position. A novel image is first transformed into its neutral form and then matched against the data base of (neutral) images.

In this work we focus on recognition and image re-rendering under lighting condition variability of a *Class* of objects, i.e., objects that belong to a general class, such as the class of faces. In other words, for the re-rendering task, given a sample images of members of a class of objects, and a *single* image of a new object of the class, we wish to render new images of the new object that simulate changing lighting conditions.

Our approach is based on a new result showing that the set of all images generated by varying lighting conditions on a collection of Lambertian objects all having the same shape but differing in their surface texture (albedo) can be characterized analytically using images of a prototype object and a (illumination invariant) "signature" image per object of the class. The Cartesian product between the signature image of an object and the linear subspace determined by the images of the prototype object generates the image space of the object. The second result is on how to obtain the signature image from a data base of example images of several objects while proving that the the signature image obtained is invariant to illumination conditions.

Our method has two advantages. First and foremost, the method works remarkably well on real

images (of faces) using a very small set of example objects — as few as two example objects. The re-rendering results are in many cases indistinguishable from the "real" thing and the recognition results outperform by far conventional methods. Second, since our approach is based on a simple and clean theoretical foundation, the limitations and breaking points can be clearly distinguished thus further increasing this algorithm's practical use.

1.1 Sources of Variability

What does a two dimensional image can reveal on a $3D$ object? The answer depends on the way the object was "grasped" by a camera, i.e. under what conditions it was taken. The variability of geometric, photometric and rigidity conditions generates an infinite possibilities for $3D$ to $2D$ mappings, and turns, what seem to be natural to human visual system, to the main obstacle of computer vision recognition tasks. Geometric source of variability can be defined as changes in the spatial location of image information as a result of relative change in viewing position. It is therefore obvious that $3D$ reconstruction of the object can diminish possible ambiguities and thus solve the one to many mapping problem [1].

Changing illumination as a source of variability drew much attention recently. The fact that the same object might appear dramatically different under varying illumination conditions such as light source intensity, direction and the number of the light sources is well demonstrated in Figure1.1 [2]. Moreover, photometric variations affect the perceived structure and texture of a $3D$ object even when it is not projected on a $2D$ image plane, in a way that might mislead an elaborate visual system such as human, that uses much more information than there is in a gray-level $2D$ matrix. It is important to note in this context the observation of Moses et al. [Adini et al., 1997] that the variability due to illumination, in face images is often greater than the variability due to a change in the person's identity.

However, there are implementations, where photometric effects serve as cues for object's shape reconstruction, using methods such as "shape from shading" described in section 2.1.

Non-rigid transformations such as facial expressions or body gestures also change object appearance. They can be ignored if the non-rigid parts are small compared to the object (such as in a task of matching a smiling face to a sad face) or if each of the moving object parts is treated separately (for example, recognizing a moving body parts such as legs and than match it to the all body).

The context (background) an object appears in, presents another source of variability. This is usually overcome by segmentation in cases the background has contextual significance. It should be noted, however, that when dealing with human faces, for example, which can appear with or without facial hair or glasses, it might be hard to determine which parts of the image are integral parts of the subject and which are removable.

Performing synthesis or analysis of images with varying viewing conditions, one faces different challenges depend on the source of variability. This work concentrates on illumination variations. Form now on we assume that the objects are fully segmented, taken from the same view point, scaling is unique and so are the images' sizes. There are no non-rigid transformation, and for the recognition part performed on human subjects we assume no glasses or facial hair.

1.2 Classes of Objects

Categorizing a set of objects can be done in totally different ways, depending on the view point and the aims of the classifier. Zoological definition for the class 'mammals' contradicts the intuitive ten-

[1] Whether $3D$ reconstruction is a necessary stage recognizing an object under geometric transformation is behind the scope of this work.

[2] Another source of variability due to illumination is the spectral content of the light, which is not demonstrated in figure1.1 and is ignored in this work.

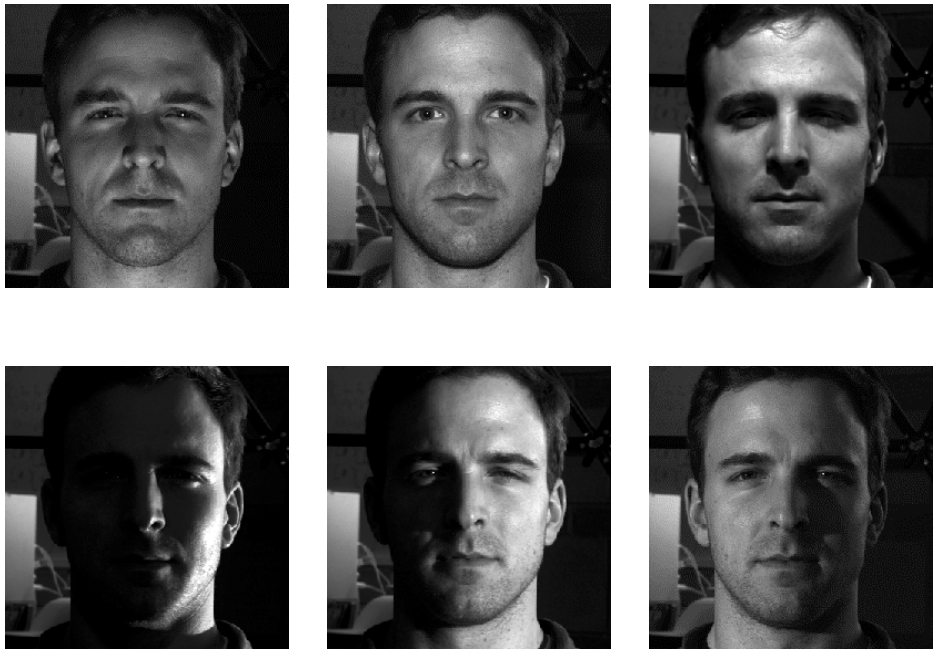


Figure 1.1. The same person seen under varying ligthing conditions can apear dramatically different. Images are taken from Belhuemer database see section 6.1. Following [Belhumeur et al., 1996] idea.

dency of a child to classify a whale, for example, as a fish.
Rosch and her colleague [Rosch, 1973, Rosch and Mervis, 1975, Rosch et al., 1976] define three levels of categories: *Superordinate* categories (e.g. animals), *Basic-Level* categories (e.g. mammals) and *Subordinate* categories (e.g. cats). Rosch et al. conducted an impressive series of studies to demonstrate the psychological reality of basic-level categories. It was found out that though classification of objects of superordinate category, necessitates high level of abstraction, objects of basic-level and subordinate categories have much more features in common (within class) and in most cases can be classified by physical cues such as shape. However, even for the well defined, basic-level class of chairs, functionality, shape and connotation do not have to be in correlation. Consider for instance a wheel chair, electric chair and a deck chair ...
Yet, when one is familiar with common features of a class, he is able to deduce, from part of the members on the others.
Since our only concern is with what can be 'seen', a class would be uniquely defined by its objects shape. Class members might vary by their reflectance (light and texture) properties, or by their pose in the scene. Many computer vision algorithms (see section 2.3), including ours, imitate human ability to learn [3], by constructing models, which, using the class-based assumption, are able to learn on a novel object of a class from a set of labeled examples.
The basic result used in this thesis is that the image space generated by varying the light source lives in a three-dimensional linear subspace [Shashua, 1992, Shashua, 1997]. Thus three images of an object, taken under different illuminations are sufficient to generate novel images of the object under any illumination conditions. The class based assumption enables us to extend this previous result and use, for the synthesis process, only one image of the object in addition to images of other objects of the class. In order to use this assumption, precise definition of a class is needed, and is given in section 3.2.

[3] See also [Tarr and Gauthier, 1998, Moses et al., 1996]

Though the discussion relates generally to any class of objects, we will refer mostly and give examples of the class of human faces– a most typical and usable example of a class.

1.3 Class of Human Faces

Machine recognition of faces from still and video images is emerging as an active research area spanning several disciplines such as image processing, pattern recognition, computer vision and neural networks. One can attribute this to the fact that, faces recognition technology (FRT) has numerous commercial and law enforcement applications. These applications range from static matching of controlled format photographs such as passport, credit cards, photo ID's, driver's license, and mug shots to real-time matching of surveillance video images presenting different constraints in terms of processing requirements. For an excellent survey on face recognition see [Chellapa et al., 1995].
Since the algorithm offered in this work is based on previously segmented and aligned images in terms of scale and pose, it is more suitable for applications based on data from still images taken under controlled conditions such as passport images. The symmetric characteristic of the "class of faces", and its comparatively small variance of head sizes and the proportion between features of the face enables a unique definition of frontal view as well as a unique determination of focal length. Thus pose and scale unity can be achieved easily over a huge set of images taken in different places. Moreover one can control these parameters quite easily, or invest little effort in a preprocessing stage needed to perform alignment.

1.4 Related Issues in Human Vision

In general, the human recognition system utilizes a broad spectrum of the senses (visual, auditory, olfactory, tactile, etc.). These stimuli are used in either an individual or collective manner for both storing and retrieval of images for the purpose of recognition. In addition there are many instances when contextual knowledge is also applied, i.e. the surroundings play an important role. However, in the research for computer algorithms, where matrices of image intensities are the only inputs, it might be interesting and even inspiring to learn on the way humans perform recognition and synthesis tasks, especially – in the scope of discussion of this thesis – under photometric variations. Many aspects of the human visual capabilities are dealt with a wide range of researches in psychophysics, neurophsychology and cognitive psychology. We will focus on two general questions. The first concerns the way visual stimuli are represented in the brain. This question has an implication on the computer vision model based approach, which recently became popular. The other issue relates to light perception, or more precisely to the impact of light on scene interpretation.

1.4.1 Visual Stimuli Representation

In a profound paper, Farah [Farah, 1988] deals the question whether visual imagery is really visual, and presents two sides of this controversial issue. One side of the debate maintains that imaging consists of the top-down activation of perceptual representations, that is, representations that are also activated automatically by an external stimulus during perception. In contrast, it is claimed that the representations used in imagery are not the representation used in perception, and that the recall of visual information, even when accompanied by the phenomenology of "seeing with the mind eye", is carried out using representations that are distinct from those used in veridical seeing. This debate is not only relevant to the question of image retrieval - or recognition - in computer vision language, but for *Mental Images*, i.e. images generated by the mind - or synthesis - in computer graphics terms. Computer vision researchers, which favor abstract or *non-pictorial* image representation on raw images, argue this preference, relying on the brain assumed behavior - which presumably uses more compact representation - without redundancies. On the other hand, it seems that cognitive psychology findings,

pioneered by Shepard (see, e.g., [Shepard and Cooper, 1982]), that shape can be mentally reoriented only with continuous "mental rotation", apparently contradict this assumption, since they provide a demonstration of the apparently visuospatial properties of mental images.
Mental rotation phenomenon, however, is given by other researchers as the analogy to image alignment, as a pre-processing stage to recognition. In this analogy, there is an implicit assumption that the brain stores a prototypical view of the retrieved object in some way, and applies on it previously learned transformations in the matching stage to a novel view. One can take this assumption even further in two directions, both have relevance to this work. The first is concerned with classes of similar (in shape) objects. Views of similar objects might look the same, under the same transformation, so it might be sufficient to learn all the possible views of an object given one or few examples of the class. The other direction can be a finding an equivalent to the mental rotation phenomena in the photometric domain. Unfortunately, no such report is known.
It should be noted that despite what might have been implied from the above discussion, similar objects might have similar representations which can be non-pictorial at all. In the same manner, views of continuous geometric transformations of an object might have representations which change successively.

1.4.2 Light Perception

One of the outstanding phenomena concerning human visual system characteristics is *Light Constancy* (for survey see [Coren and Ward, 1989]), that is the perception of object's lightness is relatively independent of the amount of light reaching one's eye (which is a product of the light source intensity and the reflectance properties of the object). For example: a piece of white paper appears to be approximately the same shade of white whether it is viewed in a dim light or bright light. A piece of coal viewed in bright sunlight will still appear black even though it may be reflecting a greater amount of light to the eye than would a white piece of paper viewed in ordinary room light. Two possible explanations were offered to the lightness constancy mechanism. Both relate to contextual knowledge. The first claims that constancy is maintained by relationship between stimuli, i.e the ratio of intensities of two patches of light on the retina is preserved. The second explanation would argue that the observer responds to cues indicating the nature of the illumination falling on the objects, and adjusts the apprehended lightness of the object in consciousness accordingly.
The light constancy phenomenon raises a more general question on light perception. Is there a primary stage where light impact is naturalized, before higher level process, such as recognition, is performed? If so, human subject should able to estimate the direction and intensity of a light source correctly. Such information can be only received from the scene. But, in order to decipher the scene, a preprocessing recognition stage might be applied, so one can claim in contradiction that recognition and light source direction recovery are done simultaneously.
An example is seen in figure 1.2a which shows thresholded face images termed after Mooney [Mooney, 1960]. It seems that in this case the illumination is factored out simultaneously with the recognition process. The thresholded images appear to be recognizable, at least in the sense that one can clearly identify the images as containing faces. Because the appearance of the thresholded images critically relies 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're looking at the image of a face) may be required in order to factor out the illumination. Notice, however, that eliminating illumination effects, which are expressed in white and black surfaces in the images (by applying level crossings on the images), as can be seen in figures 1.2b, makes recognition impossible.

A well known example which might support the opposite approach can be seen in figure 1.3. Interpretation of the image might change when the image is turned upside down, then the two lava cones with craters will be perceived as two craters with mounds. This reversal of the perceived concavities

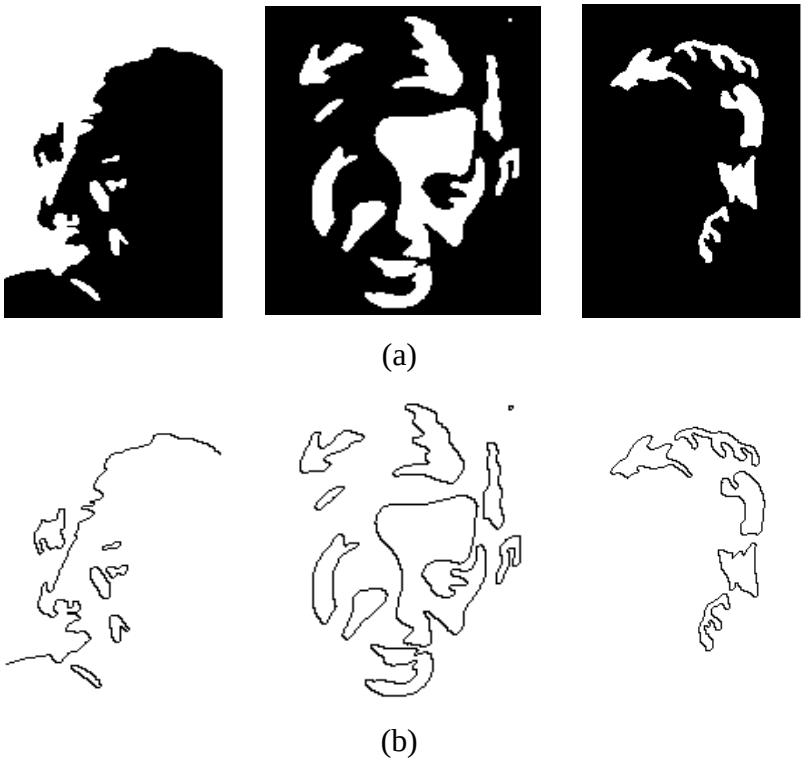


Figure 1.2. Mooney faces (a) and their level-crossings (b).

and convexities is apparently due to an implicit assumption by the viewer, that the scene is lit from above. This example suggests preliminary estimation of the light direction before scene interpretation. Support to this argument can be demonstrated using Horn's famous example [Horn, 1986] on the effect of shading on the perception of the shape of a surface, as seen in figure 1.4. The photographs demonstrate how the skillful application of makeup does more than just alter surface texture: It creates highlights and shadows that manipulate our perception of surface shape. Once again the scene is interpreted subsequently to former assumption about the effects of the light.
These two examples test human ability to decompose structure, texture and light direction given one image, as in the case of the first example. Note that even when two images of the same subject are given, as is demonstrated in the second example our visual system fails to detect makeup as simply a texture variation. The direct relevance of this observation to the Q-image algorithm will become clearer in Chapter 5.

1.5 Structure of Thesis

The next Chapter surveys briefly previously and commonly used approaches of handling recognition and synthesis tasks. The emphasis is on the Model based approach algorithms which enjoy an increased popularity recently. Basic issues concerned with the Model Based approach and Image Representation are also dealt with, due to their relevance to the thesis.
Background on the light model is given in Chapter III. Based on the definition of class based objects the Synthesis and Recognition tasks are re-defined.
Chapter IV presents an alternative approach to the Quotient Image method – the Reconstructionist Approach which is based on linear combination of the bootstrap set images, and discusses its shortcomings.

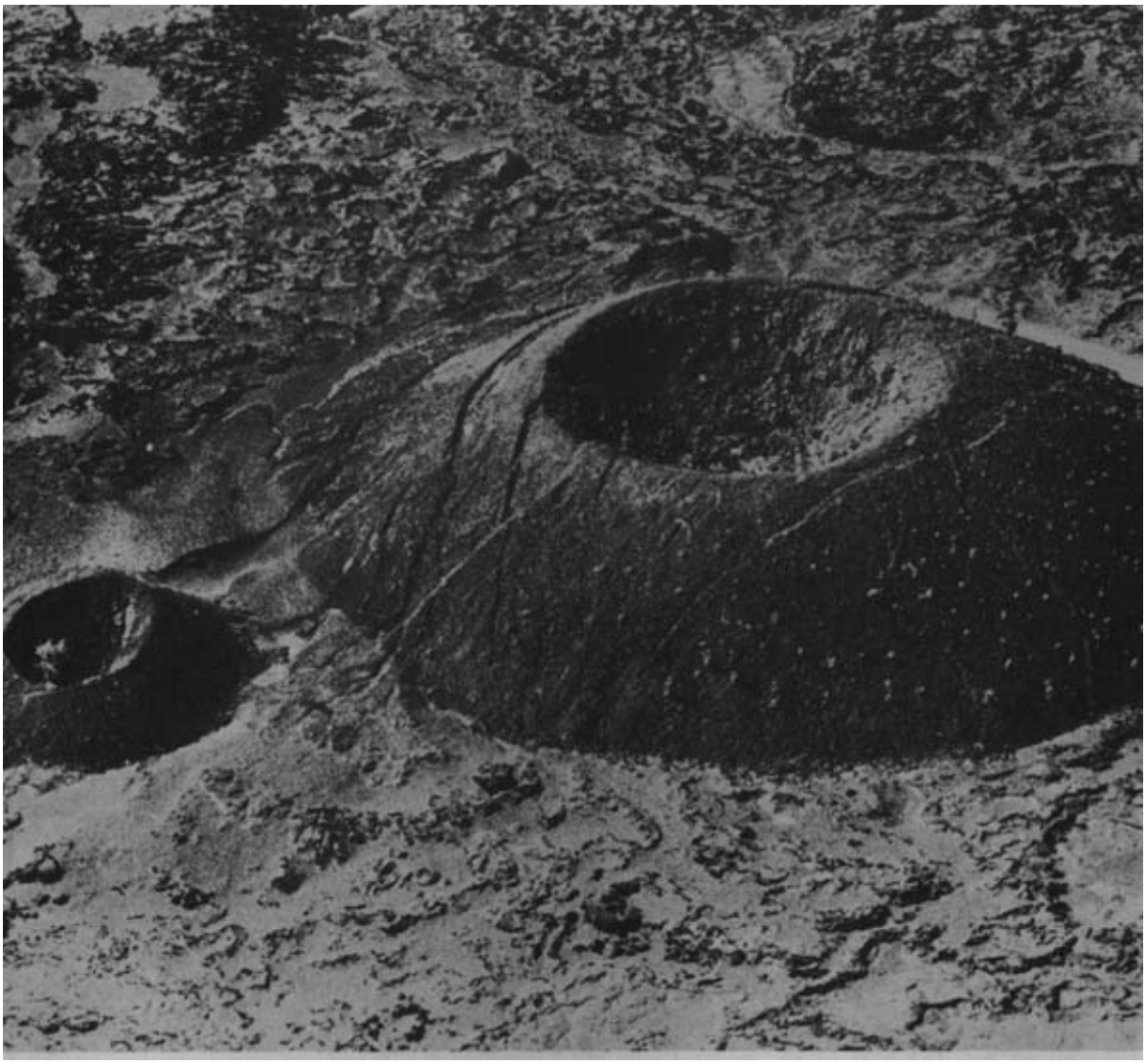


Figure 1.3. The two lava cones with craters, are percived as two craters with mounds, when the image is turned upside down. This reversal of the perceived concavities and convexities is apparently due to an implicit assumption of the viewer that the scene is lit from above. (Photograph provided by Associated Press/ Wide World Photos, 1972. Taken from [Nalwa, 1993], after [Rittenhouse, 1786]).

(a) (b)

Figure 1.4. Photograph of a model with (b) or without (a) makeup illustrating how the shading of a surface may dramatically affect out perception of the surface's shape. (Photographs, courtesy Merle Norman Cosmetics, Bellanca Avenue, Los Angeles. Taken from [Nalwa, 1993] after [Horn, 1986]).

The heart of the thesis is given in Chapter V, which describes and proves the Quotient Image algorithm. The chapter starts with a definition of an illumination invariance measure termed the Quotient Image. The Cartesian product between the quotient image of an object $\mathbf{y}$ and the linear subspace determined by the images of the prototype object generates the image space of $\mathbf{y}$ (Proposition 1). The second result is on how to obtain the signature image from a data base of example images of several objects while proving that the the signature image obtained is invariant to illumination conditions (Theorems 1,2). Description on how to extend the algorithm to handle color images is also given.
Chapter VI discusses and implements the algorithm described in Chapter V, supplying empirical proofs to the theorems and prepositions previously stated. Synthesis results of a varied collections of images are presented, for both gray-scale and colored images. Next section, goes back to the Reconstructionist Approach, demonstrating visually its lack of invariance to illumination, as well as low quality synthesis results compared to those one can get using the Quotient Image approach. The last section in this Chapter presents the recognition algorithm. Its performance is tested on a database consists of 1800 images, and is compared to other methods.
Chapter VII sums up the work and suggests some possible directions for future research.

Chapter 2

Related Approaches and Aspects of Representation

Researchers, in both computer graphic and computer vision fields have used intermediate, physically based models to approach their respective problems of synthesis and recognition. In computer graphics, sophisticated three-dimensional modeling and rendering techniques have been developed that effectively simulates the physics of rigid and non-rigid solid objects and the physics of imaging. Work in computer vision has followed a parallel path; Most object recognition algorithms used $3D$ object models and exploited the properties of geometrical and physical optics to match images to the database of models. A different approach was to extract invariant or semi-invariant measures in the object view. Early work in both approaches deal with visual recognition in the context of low level processes such as edge detection, lightness, color constancy, and shape from shading.
Recently *Model Based* approaches have been used, considering images as 'mathematical entities', such as vectors in a space. The idea, borrowed from work which handled recognition under geometric variations, gave rise to a huge class of algorithms (some of them rely on the linear behavior of light reflectance for certain surfaces and therefore can be applied only for photometric issues). In most of these algorithms, the image to be analyzed (or its representation) is compared directly, to a set of example images (or their representation).
The model based approach can be further divided into *Object Based* and *Class Based* methods. Object based methods use images of the same object under different viewing conditions [Shashua, 1992, Belhumeur et al., 1996, Hallinan, 1994], whereas class based methods use images of different objects of the same general class. Such class-based methods have the advantage that they can deal with novel object within a given class.

2.1 Classic Approaches

The traditional approach has been to recover geometric features, such as lines, curves and vertices, to hypothesize and verify the three dimensional object's structure, while directing little effort toward the explicit use of other scene properties such as reflectance, roughness, and material type. The main drawback of this approach, is the variance of features, such as edges, under varying illumination conditions. Optionally, an algorithm, which does rely on photometric properties of the surface, to recover the geometric structure of the scene, was offered, in 1977, by Horn and his colleagues [Horn, 1977]. Horn's algorithm followed by a class of works [Ikeuchi and Horn, 1981, Horn and Brooks, 1986, Pentland, 1982, Pentland, 1984] known in computer vision as *Shape From Shading* (SFS) algorithms uses gray-level values in the image to recover the shape. These algorithms failed to recover shape of non-synthetic or complex objects since they were limited by a priori informa-

tion and assumptions, one has to presume on the scene. These often include surface orientation along surface boundary, and the assumption of uniform albedo. Woodham's *Photometric Stereo* algorithm [Woodham, 1980] overcomes this obstacle, using 3 differently illuminated images, of the same scene. However, the necessity to recover the light source directions, applying SFS and Photometric Stereo makes them both not feasible.

2.2 Model Based Approach

2.2.1 Fundamental Issues

Most of the model based algorithms embed the vectorized images in a linear subspace. Doing that, they should have to care for correspondence and dimensionality, both are non trivial issues as will be described below.

Correspondence

The key underlying the mathematical assumption of the model based approach is that the images form a linear vector space. However, images are just arrays of numbers or pixels, not vectors. A set of raw images – say of similar objects – does not have the structure of a vector space, because operations like addition or multiplication by a scalar do not have a well defined meaning for raw images. In pattern recognition a standard technique for associating a vector to an image is to derive the vector components from an ordered set of measurements on the image. This technique, however, is incompatible with image-based approaches, where vector components must correspond to pixels. A vector space structure implies that the i-th component of all the vectors in the set must refer to the same type of feature on the imaged surface. Strictly speaking, the use of vector space techniques in image based approaches requires the solution of the correspondence problem: finding pixels in two or more images that represent corresponding surface features in the scene. Correspondence is a difficult problem in computer vision. It is usually solved, for sufficiently similar images, using optical flow algorithms, which find corresponding pixels in two or more images and compute their displacement vectors (in the image plane). Correspondence transforms images into vectors associating to feature point i its color (or gray level value) and its (x, y) position.
Since the work considers only photometric variation (not geometric), we can assume that given images of the same object, under different lighting conditions, the images' pixels naturally match and no correspondence process should take place – a straightforward image vectorization is sufficient [1]. As for images of different objects of the class, since the objects are similar but not identical, an alignment process is needed. We've found out that subsequent to compensating for scale and geometric transformations between the images, a center-of-mass alignment is sufficient, so that the main features of the object should "roughly" correspond as described in section 6.1.2. Notice, that if the source of variation is geometric, image correspondence can not be based on absolute position in the image but on other cues such as color or gray-level values, surrounding etc.... In these cases, using the class-based assumption, a dense correspondence is needed [Beymer and Poggio, 1996].

Dimensionality

Embedding images in a linear vector space, one should take into consideration its dimensionality. It seems that under certain circumstances the image space generated by varying the parameters of the source can be represented as a function of a small number of sample images from the image space.

[1] In case there are variations between the images, other than photometric, a preprocessing stage such as scaling, rotation, translation is needed

Image representation in the reduced image space, also referred to as *Feature Space*, should preserve the algebraic attributes of the image, being intrinsic but not necessarily visible. In this sense one no longer deals with an image but with a model of an image and this vectorization is the main principle of all the model based approaches. Image analysis, and object recognition in particular, can benefit from the reduced dimensionality, as long as the characterized object information is preserved, not only in terms of computational time savings. The key idea of most of the dimensionality reduction techniques is to factor out the non-relevant and the misleading information, such that the source of variability, which generates differences between images of the same object, would hardly effect the recognition process.

As for image synthesis, the low dimensionality of the image space under lighting variations is useful for rendering novel images given a small number of model images, or in other words,it provides the means for an *Image-Based Rendering* process in which sampled images replace geometric entities formed by textured micro-polygons for rendering new images. Dimensionality reduction of the image space is handled in various ways in most of the recent works. Some, are unique to computer vision, while others are borrowed from other domains: Karhunen-Loeve transform, Singular value decomposition, Neural network classifier, Fisher discriminant analysis and Support Vector Machines. The following subsection deals these methods and their applications.

2.2.2 Current methods and approaches

Eigenpictures

The optimal way to achieve data dimensionality reduction, is what is known as *Karhuenen-Loeve* (KL) expansion in pattern recognition [Fukunaga, 1989, Ash and Gardner, 1975, Devijver and Kittler, 1982] and as *Factor* or *Principle Component Analysis* (PCA) in the statistical literature [Ahmed and Goldstein, 1975]. The KL expansion has originally been studied for image compression, but though optimal, faster transforms such as discrete sine and cosine transform have been preferred. Extensive work has been also done in the analysis of signals in the time domain. The pioneer works of Sirovich and Kirby [Sirovich and Kirby, 1987] and Turk and Pentland [Turk and Pentland, 1991] which used PCA for image representation and recognition, became a corner stone for much of the recent work done in these fields.
The main concept is as following: Let $\mathcal{P} = span\{\psi_i\}$, $i = 1 \dots N$, be an N dimensional image space where ψ_i is a vector representation of the i-th image. $\mathcal{P}$ can be also spanned by a smaller set of orthogonal vectors $\{u_j\}$ $(j <= i)$ Let Ψ be a matrix which its columns are ψ_i. One of the ways to obtain the $\{u_j\}$ is to decompose the matrix $(\Psi - \hat{\Psi})^T(\Psi - \hat{\Psi})$ by SVD, where Ψ^T is the transpose of Ψ and $\hat{\Psi}$ is the average of the columns of Ψ. SVD can be viewed as a deterministic counterpart of the KL transform. The singular values (SV's) of an image are very stable and $\{u_j\}$ are termed *Eigenvectors* or *Eigenpictures* in computer vision context. One can think of this process as a rotation of the referred coordinate system so that as few axes as possible will convey most of the information. If the images' points are not spanned uniformly in the N dimensional image space (as is the case in similar images - such as images of faces), dimension can be reduced by neglecting the less informative axes (axes on which the projected data points have the lowest variance). The data points variances along the eigenvectors are indicated by the corresponding eigenvalues, thus the number of significant eigenvectors to be selected can be determined by applying a threshold on the eigenvalues. Usually, it can be done quite easily since these values tend to descend in a step-like manner. Once the eigen vectors are obtained, any image in the ensemble can be approximately reconstructed using a weighted combination of the eigen vectors. The weights that characterize the expansion of the given image in terms of eigen pictures serve the roll of features.

Neural Networks Classifiers

One of the commonly used tools to map an image into a feature vector for analysis purposes, or to do the inverse mapping for synthesizing new images, are *Neural Networks* (NN). The network, usually trained on a set of labeled examples, should be able to generalize, and label correctly future set of (as wide as possible) similar inputs. Since this field, though popular, has no relevance to the thesis we will only mention few, not necessarily representative, works have been done in the past years.
One of the earliest reports for the use of NN for face recognition was reported in [Kohonen, 1988] and termed the *Kohonen Associative Map*.
Oja [Oja, 1992, Oja, 1995] introduced a neural network architecture that provides a novel way for parallel on-line computation of PCA expansion.
Recently, a new wave of algorithms [Foldiak, 1990, Field, 1994, Olshausen and Field, 1995] and many others, based on novel approaches, such as *Sparse Coding* or *Independent Component Analysis* (ICA) seem to replace the extensive use of "linear" PCA for recognition purpose. These techniques are non-linear, and the only applications known to generate sparse coding use neural networks, (see for example [Meunier and Nadal, 1995]).

Linear Combinations of Models

In a seminal paper, Ullman and Basri [Ulman and Basri, 1991] showed that for orthographic projection, the set of all possible images of an object undergoing rigid $3D$ transformations and scaling is embedded in a linear space and spanned by a small number of $2D$ images. They have proved that only three views are needed for general rotation and rigid transformation and scaling in the $3D$ space, whereas for linear transformations two views suffice. However, not the images but the images' points were used in their proofs, and the results were demonstrated on *silhouettes*– images generated by the orthographic projection of the objects rims [2]. Since this work, done in geometric domain, exceeds the main theme of the thesis, we will only give a brief review of some the main principles, which inspired the research in the photometric domain[3], to be described in the next subsection (Section 2.2.2). The main claim deals with images of an object undergoing a linear transformation in $3D$ space. Let O be a set of object points. Let P_1, P_2, P_3 be three images of O obtained by applying 3×3 matrices R, S and T to O respectively. (In particular R can be the identity matrix, and S, T can be two rotations producing the second and the third views). Let $\hat{P}$ be the fourth image of the same object obtained by applying a different 3×3 matrix U to O. Let r_1, s_1, t_1 and u_1 be the first row vectors of R, S, T and U, respectively, and let r_2, s_2, t_2 and u_2 be their second row vectors. The positions of a point $p \in O$ in the four images are given by:

$$p_1 = (x_1, y_1) = (r_1 p, r_2 p)$$
$$p_2 = (x_2, y_2) = (s_1 p, s_2 p)$$
$$p_3 = (x_3, y_3) = (t_1 p, t_2 p)$$
$$\hat{p} = (\hat{x}, \hat{y}) = (u_1 p, u_2 p)$$

The claim is that if both sets r_1, s_1, t_1 and r_2, s_2, t_2 are linearly independent, then there exist scalars a_1, a_2, a_3 and b_1, b_2, b_3 such that for every point $p \in O$, it holds that

$$\hat{x} = a_1 x_1 + a_2 x_2 + a_3 x_3$$

[2] *Rim* is the set of all the points on the object surface whose normal is perpendicular to the viewing direction

[3] Shashua, first presenting his novel photometric alignment model [Shashua, 1992] claimed to do a similar use of the linear combination approach as appeared in the work of Ullman and Basri [Ulman and Basri, 1991].

$$\hat{y} = b_1 y_1 + b_2 y_2 + b_3 y_3$$

The proof is derived immediately. Since the two sets are linearly independent each spans $\mathcal{R}^3$, u_1 and u_2 can be expressed as linear combinations of r_1, s_1, t_1 and r_2, s_2, t_2 respectively. $\hat{x} = u_1 p$ and $\hat{y} = u_2 p$ and that completes the proof.
This result was farther extended to handle general rotation and rigid transformations and scaling in the $3D$ space. In addition the authors show how two views can suffice in general linear transformations. For wider scope see [Ulman and Basri, 1991].
The Tomasi and Kanade rank theorem [Tomasi and Kanade, 1992], termed the *Factorization Method*, presents a different approach to the same idea. An extension to the perspective case can be found in [Shashua, 1995].

Photometric Alignment

The basic result about the low dimensionality of the image space under varying lighting conditions was originally reported in [Shashua, 1992, Shashua, 1997] in the case of Lambertian objects. Shashua showed that an image of an object can be represented as a linear combination of a fixed set of k images of the object. Moreover, $k = 3$ in case the surface is matte – this observation was made independently by Yael Moses. The proof can be stated as following: Let $I(p)$ be the gray-value of pixel p in image I. It can be represented as

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

Here, the length of the surface normal $\mathbf{n}_p$ represents the surface albedo,(a scalar ranging from zero to one). The length of the light source vector 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 the image in the surface. Now, let $\mathbf{a}_1, \ldots, \mathbf{a}_k$ be some arbitrary set of basis vectors that span the k-dimensional Euclidian space. The image intensity $I(p) = \mathbf{x}(p) \cdot \mathbf{a}$ is therefore represented by

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

where $\alpha_1 \ldots \alpha_k$ are the linear coefficients that represent $\mathbf{a}$ with respect to the basis vectors, and $I_1 \ldots I_k$ are the k images $I_k(p) = \mathbf{x}(p) \cdot \mathbf{a}_k$.
Alignment based recognition under changing illumination can proceed in the following way. The images $I_1, \ldots, I_k$ are the model images of the object (three for Lambertian under point light sources). For any new input image I, rather then 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 than synthesize an image $I' = \alpha_1 I_1 \ldots \alpha_k I_k$. If the image I is of the same object and the only change is in illumination, then I and I' should perfectly match. 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 an experiment conducted by Hallinan [Hallinan, 1994], images of a face viewed from a fixed direction and different illumination, were analyzed using the PCA approach. The firsts eigen faces show indubitably the principal light source direction as is demonstrated in figure 2.1 generated by us following Hallinan. Based on Shashua's result, under the assumption of linearity, the first three span the image space. Expanding the model to handle non-Lambertian and self shadowing surfaces, Hallinan claims that the first five eigenfaces can consistently be interpreted as representing five very different lighting situations. He bases this conclusion from inspection of several examples and on his ability to synthesize new images of the analyzed face using five eigenfaces with satisfying similarity to the

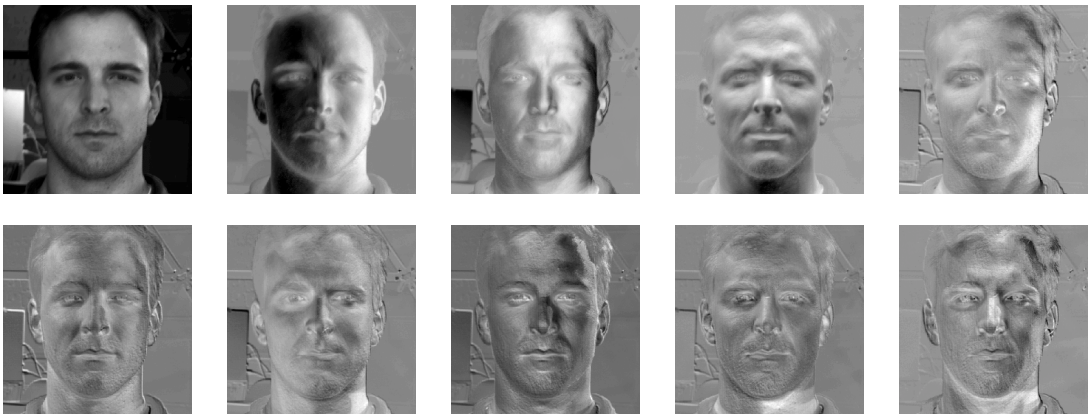


Figure 2.1. The first 10 eigenfaces of one subject, generated out of 47 differenly illuminated images taken from Belhuemer database. The first eigenface is what we term "DC" image (or avarage), which looks very similar to the original images. The next three convey most of the information on illumination directions.

original [4].

In one of his earliest papers concerning face recognition under varying illumination, Kriegman and his colleagues [Belhumeur et al., 1996], offered and examined several algorithms based on the photometric alignment approach. The key idea is to generate an object's representation (face, in the referred paper) using three images of the object under three distinct light conditions without self-shadowing. Once, each of the objects in the database is represented as a mathematical entity (a vector or a space), new images representations are compared to the database. The use of three images that span the object's subspace (under any illumination) discounts lighting effects from the representation. Follows, a brief discussion of each of the algorithms.

One of the ideas is to do matching based on the distances between the $3D$ subspaces, spanned by the images of each object.

Another algorithm, which was conducted with some variation by us, for purpose of comparison [5], is based on generating a set of eigenfaces from the images' database. Each image can be than reconstructed using a linear combination of the eigenfaces [6]. The set of the eigenvectors' coefficients (or weights) defines uniquely each image and will be termed in this discussion *Feature Vector*. The next step will be to average each set of three feature vectors belong to each of the objects in the database. This averaged feature vector will be consider as the "object representation". Now, if the three views of each object, are of the same three lights for all the objects, the averaging step, will average the weights of the eigenvectors which donate mostly to the illumination components in the images, and thus reduces the variability due to light between the objects representations. Moreover it will enhance the typical features of each object. Given a novel image, it can be represented in the same manner as a set of eigenvectors' weights (feature vector), the distance of this vector to each of the objects representations

[4] We doubt Hallinan's observation, since it is not reasonable that non-linear illumination conditions can be precisely described by using linear method such as PCA. Following Halinan we've conducted similar experiment, and based our assumption on the eigenvalues behavior. As mentioned above the image space dimensionality is determined by the manner the eigenvalue descend as shown in figure 2.2

[5] more details and results can be found in the appendix, section **??**

[6] Notice, that not as Helinan's suggestion, the PCA is applied here on a matrix consists of images of different faces, taken under three illumination conditions

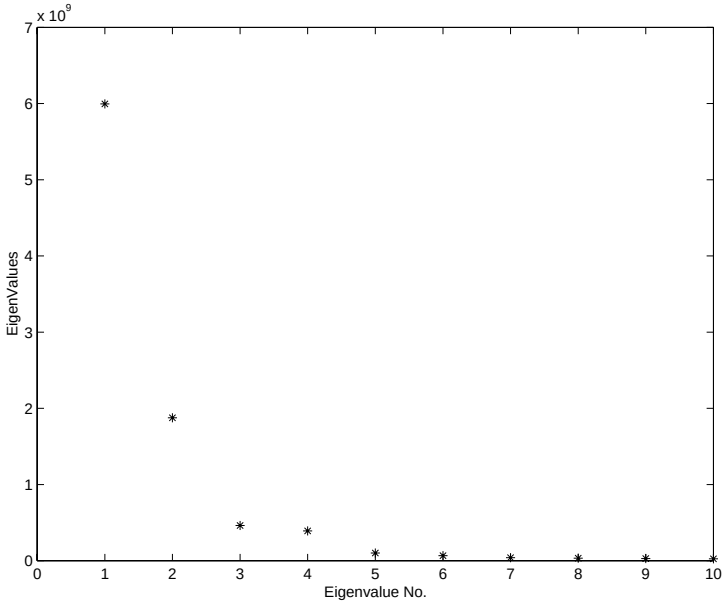


Figure 2.2. The coresponding eigenvalues to the above eigenfaces, shown in Fig. **??**. Only the first four eigenvalues are significant with respct to the rest. As seen in the upper figure, the first four eigenfaces convey most of the information and should be sufficient to span the image space of all the images of this face under varying illumination. The fact that there are four instead of the three expected from Shashua results [Shashua, 1992], can be explained by he existence of non-linearities such as highlights and cast shadows.

in the database was calculated using LSE (least square error). The shortest distance might indicate on the best match.
The discussed Kriegman's paper claims for the algorithm that is based on *Fisher's Discriminant Analysis*. As in eigenfaces methods, a set of basis vectors W – termed *Fisher-Faces*– which span the image space, is generated. The method selects W in such a way that the ratio of the between-class scatter and the within-class scatter is maximized [7]. Let the between class scatter matrix be defined as

$$S_B = \sum_{i=1}^{c} |\chi_i|(\mu_i - \mu)(\mu_i - \mu)^T$$

and the within-class scatter matrix be defined as

$$S_W = \sum_{i=1}^{c} \sum_{\chi_k \in \chi_i} (\chi_k - \mu_i)(\chi_k - \mu_i)^T$$

where c is the number of classes, i.e. the number of objects, μ_i is the mean image of class χ_i, and $|\chi_i|$ is the number of samples in class χ_i. If S_W is nonsingular, the optimal projection W_{opt} is chosen as that which maximizes the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples, i.e.

$$W_{opt} = argmax_W \frac{|W^T S_B W|}{|W^T S_W W|} = [w_1 w_2 \dots w_m]$$

where $\{w_i | i = 1, 2, \dots, m\}$ is the set of generalized eigenvectors of S_B and S_W corresponding to set of decreasing generalized eigenvalues $\{\lambda_i | i = 1, 2, \dots, m\}$, i.e. $S_B w_i = \lambda_i S_W w_i$, $i = 1, 2, \dots, m$. An upper bound on m is $c - 1$. Now, each image is represented as a set of coordinates (a point in the image space) where $\{w_i\}$ are the axes. The matching is done according to the distance between the points.
These algorithms, though none of them was expanded or currently used, formed the base to a later work of Yale's group, which displays a more elaborate object representation termed the *Illumination Cone*. The illumination cone is claimed to model the complete set of images of an object with lambertian reflectance under an arbitrary combination of point light source at infinity, using a small set of training images (see [Belhumer and Kriegman, 1997] and [Georghiades et al., 1998] for extension to non-convex objects). The model, fundamentally leaned on the linearity of the space of images of convex and lambertian objects, taken under different lights, uses integrability constraint to reconstruct surface of the $3D$ object. Shadows cast by the non-convex parts of the object can then be recovered. Recognition step can be applied by matching a given novel image to the object with the closest illumination cone.

first, an image I taken under point light source s is of the form

$$I = max(Bs, 0)$$

where B is the product of the surface normal and the albedo. $max(., 0)$ sets to zero all the negative components of I. The pixels which had been set to zero correspond to the surface points lying on attached shadows, (the convexity assumption enables to avoid dealing with cast shadows in this stage). However elegant this solution might be, the fact that shadowed areas in the image convey no information cannot be overcome. Thus, partly shadowed three images might not be sufficient to span an image space. Instead, applying SVD on a matrix consisting of much more than three images and taking three, most significant, basis vectors denoted by B^* is suggested. Moreover, using B^* instead of B assures robustness. It has been shown in [Belhumeur et al., 1997, Yuille and Snow, 1997] cited in

[7] Since the images are labeled, class is a set of views of the same object

[Georghiades et al., 1998] that from multiple images where the light source direction is unknown, a Lambertian surface can be recovered up to a family of three parameters given by a *Generalized Bas-Relief* (GBR) transformation. This family scales the relief (flatten or extrudes) and introduce an additive plane. Consequently the light source direction differs from the true light by a GBR transformation. Shadows are preserved under this transformation, thus reconstructing the extreme rays of the cone (images), the authors first reconstruct the surface (the height function) and then use ray-tracing techniques to determine which points lie on cast shadows. It should be noted that the vector field B^* estimated via SVD may not be integrable and so prior to reconstructing the surface up to GBR, integrability of B^* is enforced. Though it claims for good results of handling shadowed images, the algorithm offered is quite complicated for implementation, and involves too much computational steps, in performing tasks otherwise can be done much simply.

Generalization Versus Factorization

Model based approach algorithms can be dichotomized into two categories, depending on the strategy adopted in the process of image vectorization. The first, which is common to most algorithms reviewed so far, is to approach the task by *Generalization*. The idea is to try to grasp the widest common denominator of the set of images, despite the variations. The model should be able to generalize for novel images of the set, while still repelling images of different, though similar, objects. The second approach, is to *Factorize* an image into its variant and non-variant components. The non-variant components should be common to all the images of the same object and thus serve for recognition tasks, while the variant components can be used for synthesis of new images, not necessarily of the same object.
Following are two more algorithms which demonstrate indubitably each of the two approaches. The quotient image method to be presented here, belongs to the category of factorization methods.

Bilinear Model

The *Bilinear Model* of Tenenbaum and Freeman [Freeman and Tenenbaum, 1997] is a representative example of factorization methods. Given a matrix consisting of several views of different objects, the offered algorithm is claimed to factorize it into feature vectors, unique to each of the objects and to those characterizing the source of variation, or, in the paper terminology, to separate “content” from “style”. As a first step decomposing such a matrix into style and content matrices, the use of SVD is suggested. (However, the SVD technique does not promise correct decoupling of style from content.) The generalization task is then to classify observation in the remaining styles (the styles which did not participate the initial training set), that is, to estimate both content labels as well as style parameters. Such a task presents a classic “chicken and egg” problem, since neither the style nor the content of the new data is known priorly. Commonly used techniques of solving this type of problems are iterative. Assuming the content, style can be approximated and vice versa. The algorithm is hoped to converge after a countable number of iterations. To simultaneously classify known content in a new style and estimate new style parameters, The *Expectation Maximization* (EM) algorithm is used, which alternates between estimating the most likely content labels given the current style parameter estimates (E-step) and estimating the most likely style parameters given current the label estimates (M-step), with likelihood determined by the *Gaussian Mixture Model* [Tenenbaum and Freeman, 1997]. If in addition the test data are not segmented according to style, style labels can be estimated simultaneously as part of E-step.

Support Vector Machines

A typical example of generalization is an interesting implementation of the *Support Vector Machines* (SVM) algorithm offered by Pontil and Verri [Pontil and Verri,] for image classification. In this application, each n-pixel image is a point in an n-dimensional image space. Given a set of points which belong to either of two classes, a SVM determines the hyperplane leaving the largest possible fraction of points on the same side, while maximizing the distance between the two classes. This hyperplane is determined by a special subset of the points of the two classes, named support vectors. More formally: For a given set S of N points $\mathbf{x}_i \in \mathcal{R}^n$

with $i = 1, 2, \ldots N$, each point $\mathbf{x}_i$ belongs to either of the two classes and thus is given a label $y_i \in \{-1, 1\}$. S is *Linearly Separable* if there exist $\mathbf{w} \in \mathcal{R}^n$ and $b \in \mathcal{R}$ such that $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) >= 1$. The pair $(\mathbf{w}, b)$ defines an hyperplane of equation $\mathbf{w} \cdot \mathbf{x} + b = 0$ named *Separating Hyperplane*. Once the separating hyperplane parameters were found any new point can be classified according to it. The novel point is a new view of the object which is characterized by one of the sides of the hyperplane.

2.3 Class Based approaches

Work on *Class-Based* synthesis and recognition of images (mostly with varying viewing positions) was reported in [Beymer and Poggio, 1996, Basri, 1996, Freeman and Tenenbaum, 1997, Vetter and Poggio, 1997, Vetter and Poggio, 1996, Vetter and Blanz, 1998, Vetter et al., 1997, Edelman, 1995, Atick et al., 1997, Rowland and Perrett, 1995]. These methods adopt a *Reconstructionist* approach (see also Chapter 4) in which a necessary condition for the process of synthesis is that the original novel image be generated, reconstructed, from the database of examples. For example, the "linear class" of [Vetter and Poggio, 1997, Poggio and Vetter, 1992] works under the assumption that 3D shapes of objects in a class are closed under linear combinations (in 3D).
Recently, [Sali and Ullman, 1998] have proposed to carry an additive error term – the difference between the novel image and the reconstructed image from the example database. During the synthesis process, the error term is modified as well, thus compensating for the difference between the image space that can be generated from the database of examples and the desired images. That is, if $\{F_i\}$ is a set of feature vectors of similar objects, F_0, a novel image of an object of the class is given by

$$F_0 = \sum_i \alpha_i F_i + \Delta$$

Where Δ is the residual error. This residual error can be large especially if a small number of example views is used, or if the object in the example views are not very similar to the approximated object. Now, let F_{approx} be the closest approximation to F_0, given from the linear combination without the additive error term, i.e. $F_{approx} = \sum_i \alpha_i F_i$. Any linear operator L applied to both sides of the equation will yield:

$$LF_{approx} = \sum_i \alpha_i (LF_i)$$

Constructing a new image of the novel object, under new viewing conditions, Δ should be transformed accordingly (since it is not invariant). In the case of changing illumination conditions Ullman and Basri suggested to use $\Delta cos(\alpha)$ as the optimal estimator for the new approximation error, where α is the angle between the old and the new illumination vectors, given that both are of the same magnitude.
The lack of invariance of the error term, lessens the algorithm feasibilty, in the sense it necessiates an exact knowledge of the illumination directions. The reconstructionist approach to be displayed in Chapter 4 overcomes this limitation, using images of three distinct illuminations, based on the photometric alignment result. Our quotient image method takes farther the general concept, looking for an illumination invariant term (signature image), instead of an error term, that makes for the difference (in

a multiplicative sense) between the image space spanned by images of objects of a class and the novel image.

2.4 Reflectance Ratio

Image intensity of a lambertian surface point I, illuminated by a point light source at infinity can be defined, in a simplified model, as a product of the surface reflectance properties ρ with the cosine angle between the normal to the surface at that point, N^T, and the light source direction $\mathbf{s}$. This relation can be expressed as follows : $I = \rho N^T \cdot \mathbf{s}$.
However, none of the components in the above expression can be decoupled, unless the others are known. Instead, given two points with two (out of the three) similar components, the ratio between the third components of the two points can be recovered.

For instance, the *Gradient Ratio Constants* proposed by Wolff et al. [Fan and Wolff, 1997, Wolff and Angelopoulou, 1994] are derived from the ratio of two corresponding points with similar albedo, and light source. These quantities are used in addition to the integrability constraint for surface curvature and shape reconstruction.
Nayar and Bolle [Nayar and Bolle, 1995] estimated the *Reflectance Ratio* for each region in an image with respect to its background. This derivation is based on the observation that neighboring points on a smoothly curved surface have similar surface normals and illumination conditions.
It should be noted that despite the similarity between this work on reflectance ratio, and our notion of quotient image, both were done independently and specify different uses of the albedo ratio constant. While Nayar and Bolle use this invariant measure to define regions within a **single** image and thus give a model of the image, which will later serve for recognition tasks, our quotient image is the ratio between two images of different objects.

Chapter 3

Background and Definitions

3.1 Light Model

[1] The reflectance of a surface depends on its roughness and material properties. In general, incident light is scattered by a surface in different directions. This distribution of reflected light can be described as a function of the angle of incidence, the angle of emittance, and the wavelength of the incident light. Consider an infinitesimal surface patch with normal n, illuminated by monochromatic light of wavelength λ from the direction s and viewed from the direction v. The reflectance of the surface element can be expressed as:

$$r(s, v, n, \lambda)$$

Now consider an image of the surface patch. If the spectral distribution of the incident light is $e(\lambda)$ and the spectral response of the sensor is $s(\lambda)$, the image brightness value produced by the sensor is:

$$I = \int s(\lambda)e(\lambda)r(s, v, n, \lambda)d\lambda$$

for the purpose of discussion, let us assume the surface patch is illuminated by "white" light and the spectral response of the sensor is constant within the visible light spectrum, then $s(\lambda) = s$ and $e(\lambda) = e$. We get: $I = s\ e\ \rho\ R(s, v, n)$ where $\rho R(s, v, n)$ is the integral of $r(s, v, n, \lambda)$ over the visible spectrum. We have decomposed the result into $R(.)$ which represents the dependence of surface reflectance on the geometry of illumination and sensing, and ρ which may be interpreted as the fraction of the incident light that is reflected in all directions by the surface. Incident light that is not reflected by the surface is absorbed and/or transmitted through the surface. Two surfaces with the same distribution function $R(.)$ can have different reflectance coefficients ρ. As a result of white-light assumption, the reflectance coefficient ρ is independent of wavelength. This enables us to represent the reflectance of the surface element with a single constant.

The same can be achieved by using an alternative approach which does not require making assumptions about the spectral distribution of the incident light and the spectral response of the sensor. Consider a narrow band filter with spectral response $f(\lambda)$, placed in front of the sensor. Image brightness is then:

$$I = \int f(\lambda)s(\lambda)e(\lambda)r(s, v, n, \lambda)d\lambda$$

Since the filer is a narrow-band filter, it essentially passes a single wavelength λ' of reflected light. Its spectral response can therefore be expressed as:

$$f(\lambda) = \delta(\lambda - \lambda')$$

[1] An extensive part of this section was quoted from [Nayar and Bolle, 1995].

The image brightness measured with such a filter is:

$$I = s'e'r(s, v, n, \lambda')$$

where $s' = s(\lambda')$ and $e' = e(\lambda')$. Once again the reflectance function can be decomposed into a scattering function and a reflectance coefficient:

$$I = s'e'\rho'R'(s, v, n)$$

In this case, R'(.) represents the distribution of reflected light for a particular wavelength of incident light. On the other hand, for white light illumination, $R(.)$ represents the distribution computed as an average over the entire visible light spectrum. However the individual terms in both expressions for white-light and narrow band filters represent similar effect. In our discussion we will use the following expression for image brightness:

$$I = \kappa\rho R(s, v, n)$$

The constant $\kappa = s \cdot e$ accounts for the brightness of the light source and the response of the sensor. The exact functional form of $R(s, v, n)$ is determined to a great extent by the microscopic structure of the surface; Generally $R(.)$ includes a diffuse component and a specular component [Nayar et al., 1991]. Once again, the reflection coefficient ρ is the fraction of the incident light that is reflected by the surface. It represents the reflective power of the surface and will be referred to in this text as *Surface Albedo*. A perfectly diffuse or *Lambertian surface*, which is an idealization of a matte surface (as opposed to glossy or specular surface), has the property that its radiance depends on the illumination, and not on the viewing direction. Each point on the surface appears equally bright from all directions. The brightness of the points depends only on the amount of light incident per unit area, which is proportional to the cosine of the incident angle for a single distant light source. In these cases $R(s, v, n) = s \cdot n$ and we get

$$I = \rho\mathbf{n} \cdot \mathbf{s}$$

where $\kappa\rho \Rightarrow \rho$, $\mathbf{n}$ is the surface normal and $\mathbf{s}$ is the point light source direction.
It is important to note that the above expression does not hold for shadowed parts of the surface, thus an extension of it, or special treatment is needed since most of the images are shadowed. In fact, only surfaces viewed from the direction of the light source appear without shadows at all. We distinguish between two types of shadows: *Attached Shadows* and *Cast Shadows*. While the latter are shadows the object casts on itself (or one element in the scene on the other) and must satisfy global conditions, attached shadows are defined by local geometric conditions. We say that a point P is in an attached shadow if the angle between the surface normal and the direction of the light source is obtuse, thus $n_p \cdot s < 0$. An object point P is in cast shadow if it is obstructed from the light by other parts of the object, that is cast shadows are typical to objects with concave parts which can cast shadows on themselves. To reduce the offered algorithm treatment to attached shadows, convexity is assumed.

3.2 Ideal Class of Objects

We define next what is meant by a "class" of objects. In order to get a precise definition on which we can base analytic methods we define what we call an "ideal" class as follows:

Definition 1 (Ideal Class of Objects) *An ideal class is a collection of 3D objects that have the same shape but differ in the surface albedo function. The image space of such a class is represented by:*

$$\rho_i(x, y)n(x, y)^T s_j$$

where $\rho_i(x, y)$ is the albedo (surface texture) of object i of the class, $n(x, y)$ is the surface normal (shape) of the object (the same for all objects of the class), and s_j is the point light source direction, which can vary arbitrarily.

In practice, objects of a class do have shape variations, although to some coarse level the shape is similar, otherwise we would not refer to them as a *Class*. The ideal class could be satisfied if we perform pixel-wise dense correspondence between images (say frontal images) of the class. The dense correspondence compensates for the shape variation and leaves only the texture variation. For example, Poggio and colleagues [Vetter et al., 1997] have adopted such an approach in which the flow field and the texture variation were estimated simultaneously during the process of synthesizing novel views from a single image and a (pixel-wise pre-aligned) data base. The question we will address during the experimental section is what is the degree of sensitivity of our approach to deviations from the ideal class assumption. Results demonstrate that one can tolerate significant shape changes without noticeable degradation in performance, or in other words, there is no need to establish any dense alignment among the images beyond alignment of center of mass and scale.

From now on when we refer to a class of objects we mean an *Ideal* class of objects as defined above. We will develop our algorithms and correctness proofs under the ideal class assumption.

3.3 Tasks Definitions

Under the restriction to objects of a class, with lambertian reflectance function, Recognition and synthesis problems are defined as follows:

Definition 2 (Recognition Problem) *Given $N \times 3$ images of N objects under 3 lighting conditions and $M \gg N$ other objects of the same class illuminated under some arbitrary light conditions (each), identify the $M + N$ objects from a single image illuminated by some novel lighting conditions.*

Note that we require a small number N of objects, 3 images per object, in order to "bootstrap" the process. We will refer to the $3N$ images as the "bootstrap set". The synthesis problem is defined similarly,

Definition 3 (Synthesis (Re-rendering) Problem) *Given $N \times 3$ images of N objects of the same class, illuminated under 3 distinct lighting conditions and a single image of a novel object of the class illuminated by some arbitrary lighting condition, synthesize new images of the object under new lighting conditions.*

To summarize up to this point, given the ideal class and the synthesis/recognition problem definitions above, our goal is: *we wish to extend the linear subspace result of [Shashua, 1997] that deals with spanning the image space $\rho n^\top s$ where only s varies, to the case where both ρ and s vary*. We will do so by showing that it is possible to map the image space of one object of the class onto any other object, via the use of an illumination invariant signature image. The recovery of the signature image requires a bootstrap set of example images, albeit a relatively small one (as small as images generated from two objects in our experiments). The remainder of the work deals with exactly this problem. We first describe a "brute-force" approach for addressing the inherent bilinearity of the problem, detailed next, and then proceed to the main body of this work.

Chapter 4

A Reconstructionist Approach and its Shortcomings

We wish to span the image space $\rho n^\top s$ where both ρ and s vary. Let s_1, s_2, s_3 be a basis of three linearly independent vectors, thus $s = \sum_j x_j s_j$ for some coefficients $x = (x_1, x_2, x_3)$. Let $\rho_1, ..., \rho_N$ be a basis for spanning all possible albedo functions of the class of objects, thus $\rho = \sum_i \alpha_i \rho_i$ for some coefficients $\alpha_1, ..., \alpha_N$. Let y_s be the image of some new object $\mathbf{y}$ of the class with albedo ρ_y and illuminated by illumination s, i.e.,

$$y_s = \rho_y n^\top s = (\sum_{i=1}^{N} \alpha_i \rho_i) n^\top (\sum_{j=1}^{3} x_j s_j).$$

Let $A_1, ..., A_N$ be $m \times 3$ matrices whose columns are the images of object i, i.e., the columns of A_i are the images $\rho_i n^\top s_1, \rho_i n^\top s_2, \rho_i n^\top s_3$. We assume that all images are of the same size and contain m pixels. We have therefore,

$$\min_{x, \alpha_i} \mid y_s - \sum_{i=1}^{N} \alpha_i A_i x \mid^2, \tag{4.1}$$

which is a bilinear problem in the $N + 3$ unknowns x, α_i (which can be determined up to a uniform scale). Clearly, if we solve for these unknowns, we can then generate the image space of object $\mathbf{y}$ from any desired illumination condition simply by keeping α_i fixed and varying x.

One way to solve for the unknowns is first to solve for the pairwise product of x and α_i, i.e., a set of $3N$ variables $z = (\alpha_1 x, ..., \alpha_N x)$. Let $A = [A_1, ..., A_N]$ be the $m \times 3N$ matrix (we assume $m >> 3N$) obtained by stacking the matrices A_i column-wise. Thus, the vector z can be obtained by the pseudo-inverse $A^\# = (A^\top A)^{-1} A^\top$ as the least-squares solution $z = A^\# y_s$. From z we can decouple x and α_i as follows. Since the system is determined up to scale, let $\sum_i \alpha_i = 1$. Then, group the entries of z into $z = (z_1, ..., z_N)$ where z_i is a vector of size three. We have,

$$x = \sum_{i=1}^{N} z_i$$

and,

$$\alpha_i = \frac{1}{3} \sum_{j=1}^{3} \frac{z_{ij}}{x_j}.$$

There are a number of observations that are worth making. First, this approach is a "reconstructionist" one in the sense that one is attempting to reconstruct the image y_s from the data set of example images, the bootstrap set (for example, [Vetter et al., 1997, Vetter and Blanz, 1998,

(a) (b) (c)

Figure 4.1. Illustration of the "reconstructionist" approach. (a) original image, (b) image reconstructed from the the bootstrap set of Fig. 6.1, and (c) image reconstructed from a larger bootstrap set of 20 objects (60 images). The reconstruction is poor in both cases. See text for further details.

Freeman and Tenenbaum, 1997]). In practice, especially when the size of the bootstrap set is relatively small, $Az \neq y_s$. Moreover, for the same reasons, the decoupling of the variables x_j and α_i from the vector z adds another source of error. Therefore, before we begin creating synthetic images (by varying x_j) we are faced with the problem of having only some approximate rendering of the original image y_s. This problem is acute for small bootstrap sets, and therefore this approach makes practical sense only for large example sets. The second point to note is that there is some lack of "elegance" (which inevitably contributes to lack of numerical stability and statistical bias due to over-fitting[1]) in blowing up the parameter space from $N + 3$ to $3N$ in order to obtain a linear least-squares solution.

We illustrate the reconstructionist approach in practice in Fig. 4.1. We use a bootstrap set of 10 objects (30 images) displayed in Fig. 6.1, and a bootstrap set of 20 objects (not displayed here). The results of reconstruction are poor for both sets, although one notices some improvement with the larger set of 20 objects. The poor reconstruction is attributed to two main sources. First, is the size of the data base. A data base of 10 (or 20) objects is apparently not sufficient for capturing the variation among objects in the class. Second, and probably a more dominant source, is the lack of dense pixel-wise alignment among the database and the novel image. Previous work by [Vetter and Poggio, 1996, Vetter and Blanz, 1998, Vetter et al., 1997] demonstrate very good results with large databases (around 100 objects) under pixel-wise alignment.

In our approach, detailed below, we achieve two major goals: first, we do not make a reconstructionist assumption and thereby tolerate small databases without pixel-wise alignment, second we solve (linearly) for a system of $N + 3$ parameters (instead of $3N$). As a byproduct of the method of optimization we obtain an intermediate image, an illumination invariant signature image, which can also be used for purposes of visual recognition.

[1]Numerical problems due to "blowing" up parameter space for purpose of linearization can be reduced by solving a *heteroscedastic* optimization problem [Meer and Leedan, 1998], which could be quite unwieldy for large systems.

Chapter 5

The Quotient Image Method

Given two objects $\mathbf{a}, \mathbf{b}$, we define the quotient image Q by the ratio of their albedo functions ρ_a/ρ_b. Clearly, Q is illumination invariant. In the absence of any direct access to the albedo functions, we show that Q can nevertheless be recovered, analytically, given a bootstrap set of images. Once Q is recovered, the entire image space (under varying lighting conditions) of object $\mathbf{a}$ can be generated by Q and three images of object $\mathbf{b}$. The details are below.

We will start with the case $N = 1$, i.e., there is a single object (3 images) in the bootstrap set. Let the albedo function of that object $\mathbf{a}$ be denoted by ρ_a, and let the three images be denoted by a_1, a_2, a_3, therefore, $a_j = \rho_a n^\top s_j$, $j = 1, 2, 3$. Let $\mathbf{y}$ be another object of the class with albedo ρ_y and let y_s be an image of $\mathbf{y}$ illuminated by some lighting condition s, i.e., $y_s = \rho_y n^\top s$. We define below an illumination invariant signature image Q_y of $\mathbf{y}$ against the bootstrap set (in this case against $\mathbf{a}$):

Definition 4 (Quotient Image) *The quotient image Q_y of object $\mathbf{y}$ against object $\mathbf{a}$ is defined by*

$$Q_y(u, v) = \frac{\rho_y(u, v)}{\rho_a(u, v)},$$

where u, v range over the image.

Thus, the image Q_y depends only on the relative surface texture information, and thus is independent of illumination. The reason we represent the relative change between objects by the ratio of surface albedos becomes clear from the proposition below:

Proposition 1 *Given three images a_1, a_2, a_3 of object $\mathbf{a}$ illuminated by any three linearly independent lighting conditions, and an image y_s of object $\mathbf{y}$ illuminated by some light source s, then there exists coefficients x_1, x_2, x_3 that satisfy,*

$$y_s = (\sum_j x_j a_j) \otimes Q_y,$$

where $\otimes$ denotes the Cartesian product (pixel by pixel multiplication). Moreover, the image space of object $\mathbf{y}$ is spanned by varying the coefficients.

Proof: Let x_j be the coefficients that satisfy $s = \sum_j x_j s_j$. The claim $y_s = (\sum_j x_j a_j) \otimes Q_y$ follows by substitution. Since s is arbitrary, the image space of object $\mathbf{y}$ under changing illumination conditions is generated by varying the coefficients x_j. □

We see that once Q_y is given, we can generate y_s (the novel image) and all other images of the image space of $\mathbf{y}$. The key is obtaining the quotient image Q_y. Given y_s, if somehow we were also given the coefficients x_j that satisfy $s = \sum_j x_j s_j$, then Q_y readily follows: $Q_y = y_s/(\sum_j x_j a_j)$, thus the key is to

obtain the correct coefficients x_j. *For that reason, and that reason only, we need the bootstrap set* — otherwise, a single object $\mathbf{a}$ would suffice (as we see above).

Let the bootstrap set of $3N$ pictures be taken from three fixed (linearly independent) light sources s_1, s_2, s_3 (the light sources are not known). Let A_i, $i = 1, ..., N$, be a matrix whose columns are the three pictures of object $\mathbf{a_i}$ with albedo function ρ_i. Thus, $A_1, ..., A_N$ represent the bootstrap set of N matrices, each is a $m \times 3$ matrix, where m is the number of pixels of the image (assuming that all images are of the same size). Let y_s be an image of some novel object $\mathbf{y}$ (not part of the bootstrap set) illuminated by some light source $s = \sum_j x_j s_j$. We wish to recover $x = (x_1, x_2, x_3)$ given the N matrices $A_1, ..., A_N$ and the vector y_s.

We define the *normalized albedo* function ρ of the bootstrap set as:

$$\rho(u, v) = \sum_{i=1}^{N} \rho_i^2(u, v)$$

which is the sum of squares of the albedos of the bootstrap set. In case where there exist coefficients $\alpha_1, ..., \alpha_N$ such that

$$\frac{\rho(u, v)}{\rho_y(u, v)} = \alpha_1\rho_1(u, v) + ... + \alpha_N\rho_N(u, v)$$

where ρ_y is the albedo of the novel object $\mathbf{y}$, we say that ρ_y is in the *Rational Span* of the bootstrap set of albedos. With these definitions we show the major result of this paper: if the albedo of the novel object is in the rational span of the bootstrap set, we describe an energy function $f(\hat{x})$ whose global minimum is at x, i.e., $x = \text{argmin } f(\hat{x})$.

Theorem 1 *The energy function*

$$f(\hat{x}) = \frac{1}{2}\sum_{i=1}^{N} \mid A_i\hat{x} - \alpha_i y_s \mid^2 \tag{5.1}$$

has a (global) minimum $\hat{x} = x$, if the albedo ρ_y of object $\mathbf{y}$ is rationally spanned by the bootstrap set, i.e., if there exist $\alpha_1, ..., \alpha_N$ such that

$$\frac{\rho}{\rho_y} = \alpha_1\rho_1 + ... + \alpha_N\rho_N$$

Proof: Let $\hat{s} = \sum_j \hat{x}_j s_j$, thus, $A_i\hat{x} = \rho_i n^\top \hat{s}$. In vectorized form:

$$A_i\hat{x} = \begin{bmatrix} \rho_{i1}n_1^\top \\ \rho_{i2}n_2^\top \\ \cdot \\ \cdot \\ \cdot \\ \rho_{im}n_m^\top \end{bmatrix} \hat{s} = W_i\hat{s}$$

where $\rho_{i1}, ..., \rho_{im}$ are the entries of ρ_i in vector format. The optimization function $f(\hat{x})$ can be rewritten as a function $g(\hat{s})$ of $\hat{s}$:

$$\begin{aligned} g(\hat{s}) &= \frac{1}{2}\sum_{i=1}^{N} \mid W_i\hat{s} - \alpha_i W_y s\mid^2 \\ &= \sum_i \frac{1}{2}\hat{s}^\top W_i^\top W_i\hat{s} + \sum_i \alpha_i \hat{s}^\top W_i^\top W_y s \\ &+ \sum_i \frac{1}{2}\alpha_i^2 s^\top W_y^\top W_y s \end{aligned}$$

where W_y is defined similarly to W_i by replacing the albedo ρ_i by ρ_y. Because the variables of optimization $\hat{x}, \hat{s}$ in $f(\hat{x})$ and in $g(\hat{s})$ are linearly related, it is sufficient to show that the global minimum of $g(\hat{s})$ is achieved when $\hat{s} = s$. We have,

$$0 = \frac{\partial g}{\partial \hat{s}} = (\sum_i W_i^\top W_i)\hat{s} - (\sum_i \alpha_i W_i^\top) W_y s.$$

Hence, we need to show that

$$\sum_i W_i^\top W_i = (\sum_i \alpha_i W_i^\top) W_y.$$

We note that,

$$W_i^\top W_i = \rho_{i1}^2 n_1 n_1^\top + ... + \rho_{im}^2 n_m n_m^\top$$

Thus, we need to show,

$$(\sum_i \rho_{i1}^2) n_1 n_1^\top + ... + (\sum_i \rho_{im}^2) n_m n_m^\top =$$
$$(\sum_i \alpha_i \rho_{i1}) \rho_{y1} n_1 n_1^\top + ... + (\sum_i \alpha_i \rho_{im}) \rho_{ym} n_m n_m^\top$$

Note that the coefficients of the left hand side are the entries of the normalized albedo ρ. Thus, we need to show that

$$\sum_{i=1}^N \rho_{ik}^2 = (\sum_{i=1}^N \alpha_i \rho_{ik}) \rho_{yk}$$

for all $k = 1, ..., m$. But this holds, by definition, because ρ_y is rationally spanned by $\rho_1, ..., \rho_N$. □

The proof above was not constructive, it only provided the existence of the solution as the global minimum of the energy function $f(\hat{x})$. Finding $\min f(\hat{x})$ is a simple technicality (a linear least-squares problem), but note that the system of equations is simplified due to substitution while decoupling the role of $\hat{x}$ and the coefficients α_i. This is shown below:

Theorem 2 *The global minima x_o of the energy function $f(\hat{x})$ is:*

$$x_o = \sum_{i=1}^N \alpha_i v_i$$

where

$$v_i = (\sum_{r=1}^N A_r^\top A_r)^{-1} A_i^\top y_s$$

and the coefficients α_i are determined up to a uniform scale as the solution of the symmetric homogeneous linear system of equations:

$$\alpha_i y_s^\top y_s - (\sum_{r=1}^N \alpha_r v_r)^\top A_i^\top y_s = 0$$

for $i = 1, ..., N$

Proof:

$$0 = \frac{\partial f}{\partial \hat{x}} = (\sum_i A_i^\top A_i)\hat{x} - (\sum_i \alpha_i A_i^\top) y_s$$

from which it follows that:

$$\hat{x} = (\sum_i A_i^\top A_i)^{-1}(\sum_i \alpha_i A_i^\top) y_s = \sum_i \alpha_i v_i.$$

We also have:

$$0 = \frac{\partial f}{\partial \alpha_i} = \alpha_i y_s^\top y_s - \hat{x}^\top A_i^\top y_s,$$

which following the substitution $\hat{x} = \sum_i \alpha_i v_i$ we obtain a homogeneous linear system for $\alpha_1, ..., \alpha_N$:

$$\alpha_i y_s^\top y_s - (\sum_r \alpha_r v_r)^\top A_i^\top y_s = 0$$

for $i = 1, ..., N$. Written explicitly,

$$\begin{array}{llll} \alpha_1(v_1^\top A_1^\top y_s - y_s^\top y_s) & +...+ & \alpha_N v_N^\top A_1^\top y_s & = 0 \\ \alpha_1 v_1^\top A_2^\top y_s & +...+ & \alpha_N v_N^\top A_2^\top y_s & = 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \alpha_1 v_1^\top A_N^\top y_s & +...+ & \alpha_N(v_N^\top A_N^\top y_s - y_s^\top y_s) & = 0 \end{array} \tag{5.2}$$

Let the estimation matrix (above) be denoted by F, we show next that F is symmetric. The entries F_{ij}, $i \neq j$, have the form:

$$F_{ij} = y_s^\top A_j (\sum_r A_r^\top A_r)^{-T} A_i^\top y_s = y_s^\top A_j B A_i^\top y_s.$$

Note that B is a symmetric matrix (inverse of a sum of symmetric matrices). Let $E_{ij} = A_j B A_i^\top$, then it is easy to notice that $E_{ji} = E_{ij}^\top$ due to the symmetric property of B. Thus, $F_{ij} = F_{ji}$ because

$$F_{ij} = y_s^\top E_{ij} y_s = (E_{ij} y_s)^\top y_s = y_s^\top E_{ij}^\top y_s = F_{ji}.$$

□

The energy function $f(\hat{x})$ in eqn. 5.1 consists of a simultaneous projection of y_s onto the subspaces spanned by the columns of A_1, columns of A_2 and so on. In addition, during the simultaneous projection there is a choice of overall scale per subspace — these choices of scale, the α_i, are directly related to the scaling of the axes represented by $\rho_1, ..., \rho_N$ such that the albedos of the bootstrap set span (rationally) the albedo of the novel object. When $N = 1$, the minimum of $f(\hat{x})$ coincides with x iff the albedo of the novel object is equal (up to scale) to the albedo of bootstrap object. The more objects in the bootstrap set the more freedom we have in representing novel objects. If the albedos of the class of objects are random signals, then at the limit a bootstrap set of m objects ($3m$ images) would be required to represent all novel objects of the class. In practice, the difference in the albedo functions do not cover a large spectrum and instead occupy a relatively small subspace of m, therefore a relatively small size $N << m$ is required, and that is tested empirically in Chapter 6.

Once the coefficients x have been recovered, the quotient image Q_y can be defined against the average object: Let $\mathcal{A}$ be a $m \times 3$ matrix defined by the average of the bootstrap set,

$$\mathcal{A} = \frac{1}{N} \sum_{i=1}^{N} A_i,$$

and then the quotient image Q_y is defined by:

$$Q_y = \frac{y_s}{\mathcal{A}x}.$$

To summarize, we describe below the algorithm for synthesizing the image space of a novel object $\mathbf{y}$, given the bootstrap set and a single image y_s of $\mathbf{y}$.

1. We are given N matrices, $A_1, ..., A_N$, where each matrix contains three images (as its columns). This is the bootstrap set. We are also given a novel image y_s (represented as a vector of size m, where m is the number of pixels in the image). For good results, make sure that the objects in the images are roughly aligned (position of center of mass and geometric scale).

2. Compute N vectors (of size 3) using the equation:

$$v_i = (\sum_{r=1}^{N} A_r^\top A_r)^{-1} A_i^\top y_s,$$

   where $i = 1, ..., N$.

3. Solve the homogeneous system of linear equations in $\alpha_1, ..., \alpha_N$ described in (5.2). Scale the solution such that $\sum_i \alpha_i = N$.

4. Compute $x = \sum_i \alpha_i v_i$.

5. Compute the quotient image $Q_y = y_s / \mathcal{A}x$, where $\mathcal{A}$ is the average of $A_1, ..., A_N$. Replace divisions by zero by small numbers.

6. The image space created by the novel object, under varying illumination, is spanned by the product of images Q_y and $\mathcal{A}z$ for all choices of z.

5.1 A Note About Color

The process described so far holds for black-and-white images, not color images. We describe a simple approach to handle color images, *while still maintaining a grey-value bootstrap set*. In other words, given a bootstrap set of grey-value images, and a color image (represented by RGB channels) y_s of a novel object, we wish to create the *color* image space of that object under varying illumination. To that end, we will make the assumption that varying illumination does not affect the saturation and hue composition of the image, only the grey-value distribution (shades of color) of the image.

Given this assumption we first must decouple the hue, saturation and grey-value (lightness) components of the image y_s from its RGB representation. This is achieved by adopting the Hue Saturation Value (HSV) color space [Smith, 1978] often used for splitting color into meaningful conceptual categories. The transformation (non-linear) from RGB to HSV and vice versa can be found, for example, in MATLAB. The HSV representation decouples the color information into three channels (images): Hue (tint, or color bias), Saturation (amount of hue present — decreasing saturation corresponds to adding white pigment to a color), and Value (the luminance, or black-and-white information; the diagonal from $(1, 1, 1)$ to $(0, 0, 0)$ of the RGB cube). Saturation can vary from a maximum corresponding to vivid color, to a minimum, which is equivalent to black-and-white image. Once the H,S, and V images are created (from the R,G,B images), the novel image we work with is simply V. The algorithm above is applied and a synthetic image V' is created (a new image of the object under some novel illumination condition). The corresponding color image is the original H,S and the new V'. Similar approaches for augmenting black-and-white images using a color prototype image can be found in [Rowland and Perrett, 1995].

This approach allows using only grey-level images in the bootstrap set, yet accommodates the synthesis of color images from a novel color input image. Fig. 6.13 display examples on synthesizing color images from a grey-value bootstrap set.

Chapter 6

Algorithm Implementation

We have conducted a wide range of experimentation on the algorithm presented above. Database images were taken from different sources. We used high and low quality images, taken under controlled or uncontrolled conditions as well as images that were downloaded from the web, in order to demonstrate the algorithm feasibility for a wide range of databases qualities. The preprocessing stage needed depends on the original image conditions. Rough alignment is sufficient in most of the cases. Experiments presented in this chapter demonstrate numerically and visually different stages and aspects of the algorithm, some of them give empirical proofs to the theorems. At the end of this chapter synthesis and recognition results are shown. Performance of alternative algorithms is presented in comparison.

6.1 Technical Considerations and Empirical Results

6.1.1 Sources of Database and General Description

Databases from three different main sources of human frontal faces were used to test the recognition and synthesis algorithms offered above. All the Bootstrap sets were taken out of these databases. A bootstrap set consists of $N \times 3$ images of N subjects (N varies from one to 20), taken under 3 distinct illumination conditions, (the **same** 3 for all the subjects). Illumination directions and intensities were not known for two out of the three databases that were used. The novel images y_s were not necessarily taken from the same source as the bootstrap set images. Description of each database source is given below.

1. **Vetter Database**, An high quality database prepared by Thomas Vetter and his associates [Vetter et al., 1997, Vetter and Blanz, 1998]. It consists of 1800 frontal gray level images (8-bit precision) of 200 faces of men and women, with no facial hair, makeup or glasses. The subjects wore tight swimming caps and were taken with a black background so no mask was needed. The images were taken under 9 illumination conditions. Intensities, directions and the number of light sources - were not known. Image size was reduced to 256 by 256 pixels. We have chosen a bootstrap set collection of 2 to 10 objects, in cases where the novel images were taken from the same source (Vetter database) and up to 20 objects bootstrap set, if the novel images were taken from other sources (downloaded from the web, or were images of people from our lab). A collection of 10 objects bootstrap set images is shown in Fig. 6.1.

2. **Kriegman Database**, The database consists of 165 gray-level images (8-bit precision) of 15 subjects from different ages and races, with different hair style, and with facial hair. The images were taken under different illumination, face expression, with and without glasses. Intensities, directions and number of light sources - were not known. Images size was reduced to 160 by 220

Figure 6.1. The bootstrap set of 10 objects from Vetter's database of 200 objects.

pixels. The bootstrap set used consists of images of 3 faces only, as can be seen in Fig. 6.12a. No mask was used to cover any part of the images.

3. **Belhumeur Database**, Gray level Images of 14 subjects from different ages and races, with different hair style and facial hair, taken under wide range of illumination directions (carefully measured and reported). Images size was reduced to 256 by 256 pixels. We have chosen a bootstrap set collection of 4 objects which can be seen in Figure 6.2. No mask was used to cover any part of the images.

6.1.2 Preprocessing Stage

Some constraints must be met before applying the proposed algorithm on any of the images.

1. **Alignment**, The images of the bootstrap set and the novel images to be tested are "roughly" aligned, which means that the center of mass was aligned and scale was corrected. Manual alignment was done based on lips, pupils and nostrils. That is, face features should be roughly in the same positions but not necessarily overlapped.

2. **Segmentation**, Segmentation is needed for images of the bootstrap set, since background objects, might be different, unaligned or simply not suitable for the lambertian model. Hence, to ensure high quality results without artifacts in the background, for images that were not taken with homogeneous background, masks with uniform color should be used. There is no necessity to segment novel images, used for synthesis, so the background can be seen in the synthesis results.

Figure 6.2. The bootstrap set of 4 objects from Belhuemer database

Notice, however that this might cause to illumination inconsistencies in the synthesized images, between the object and its background. See for example the images at the first row of Fig 6.13.

3. **Hair Style, Facial Hair, Makeup and Glasses**, All are distinguishing features, yet, not prominent and thus must not practice as key features, performing recognition tasks. In order to demonstrate recognition results, we used Vetter Database, in which, as was already mentioned, the models had no facial hair, glasses or makeup and all wore tight swimming caps. Performing synthesis tasks, it is recommended to mask hair (if possible) in the bootstrap set (we did not do that!), for the same reasons used to justify segmentation. Facial hair, makeup and glasses can not be segmented but as shown in Fig 6.12 there is no significant effect on the quality of the synthesis results.

4. **Gray Level**, Colors were not used as cues for recognition though they could be. Synthesis tasks can be performed for color test images even if the bootstrap set is gray level, as is explained in Section 5.1 and is demonstrated in Fig 6.13.

6.1.3 Quotient Image Generation

The "signature image" is the ratio of two images, however it is not actually an image since its gray level values can range from zero to infinity. The most substantial problem, as the reader might guess, is division by zero. The straightforward way to overcome it is to substitute each value below a certain positive thresholded at the denominator, by the threshold. (say 1, if the pixels intensities range from zero to 255). To preserve consistency and get accurate recovery and synthesis results, it is recommended to replace all the "small values" in the bootstrap set images with the threshold, and use the "new" images throughout the entire algorithm. If the chosen threshold is below 1, or if the sum of the light coefficients is less than 1, the resultant quotient image should be threshold at the highest end, and then rescaled to the limited range for display purposes. One should always use the original Q-image for calculations.
Nayar and Bolle [Nayar and Bolle, 1995] which also dealt with intensities ratio (see section 2.4) suggested using the following equation to calculate the ratio between pixels intensities:

$$p(x,y) = \frac{I_1(x,y) - I_2(x,y)}{I_1(x,y) + I_2(x,y)} = \frac{\rho_1(x,y) - \rho_2(x,y)}{\rho_1(x,y) + \rho_2(x,y)}$$

instead of

$$q(x,y) = \frac{I_1(x,y)}{I_2(x,y)}$$

(where $I_1(x,y)$, $I_2(x,y)$ are intensities at (x,y), and $\rho_1(x,y)$,$\rho_2(x,y)$ are the corresponding albedos) to avoid division of non-zero values by zero. However, their proposal can not be accomplished in our synthesis process since in reconstruction of $I_1(x,y)$ given $p(x,y)$ and $I_2(x,y)$ one has to solve the following equation:

$$I_1(x,y) = \frac{I_2(x,y)(1+p(x,y))}{1-p(x,y)}$$

Now if $I_2(x,y) = 0$ then $p(x,y) = 1$ and the denominator of the reconstruction expression is zero.

6.1.4 Invariance of the Quotient Image

Our first test, shown in Fig 6.3, was to empirically verify that the quotient image is indeed invariant to illumination changes. The Q-images where thresholded (above one standard deviation) for display purposes. One can see that a bootstrap set of 10 objects yields a fairly invariant quotient image in spite of the large variation in the illumination of the novel images tested. The Q-images should also be invariant to the choice of the light sources s_1, s_2, s_3 used in the bootstrap set. This is demonstrated in Fig. 6.5 where the quotient image was generated against different choices of s_1, s_2, s_3 for the bootstrap object set (Vetter's database includes 9 images per object thus enabling us to imperiment with various bootstrap sets of the same 10 objects). Note that the novel image that was tested was not part of Vetter's database but an image of one of our lab members.

6.1.5 Accuracy of the Light Coefficients

Theorem 1 defines the condition needed for accurate recovery of the coefficients vector x – the albedo of the novel object ρ_y should be rationally spanned by the bootstrap set. However, selecting the bootstrap set images one can never assure in advance that the chosen images would satisfy the above condition. Neither a "magic number" to determine the bootstrap set size, nor a measure of quality of the images to be selected, exist. Consider for example a bootstrap set which consists of 3 images of the same object as the novel object. The albedo ρ_y is rationally spanned by the bootstrap set albedo since it is the same albedo. In this case, recovery of the x would be precise. Now, addition of more objects to the bootstrap set would lessen the accuracy of the coefficients calculated, since the influence of the images of the non-identical objects cannot be set to zero.
Since a selection in advance of the most similar images, is not feasible we have adopted the following rule of thumb: we assume that it is more probable that as the bootstrap set size increases, the probability that it would rationally span the albedo of the novel image would, in most of the cases, increase (note, that the example above is an exeption). The next imperiment was designed to test this.
The accuracy of the coefficient vector x is measured by the invariance of the quotient image against varying illumination, hence Fig. 6.4 displays Q-images generated by various bootstrap sets, as follows: we have tested the case where $N = 1$, i.e., a bootstrap set of a single object (row b), compared to a bootstrap set of $N = 10$ but where the reference object is the same object used in case $N = 1$ (instead of the average object), shown in row (f). Therefore, the difference between rows (c) and (f) is solely due to the effect of Theorem 1 on computing the coefficient vector x.
In order to rule out any special influence the average object has on the process (recall that once x has been recovered it was suggested to use the average object ψ as the reference object for the quotient image) we have also tested the case $N = 1$ where the images were deliberately blurred (to simulate an average object), yet the Q-images (row d) have not improved (compared to row c).

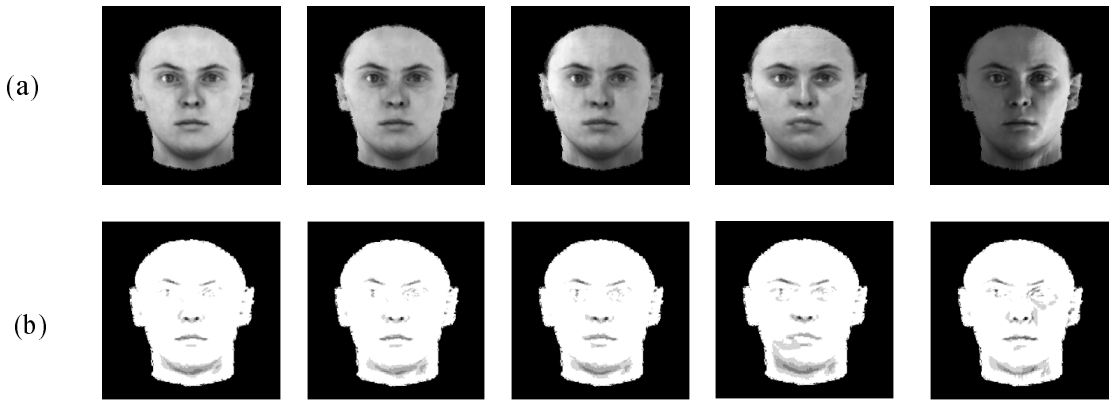


Figure 6.3. Testing the invariance of the quotient image to varying illumination. (a) Original images of a novel face taken under 5 different illuminations. (b) The Q-images corresponding to the novel images above computed with respect to the bootstrap set of Fig. 6.1.

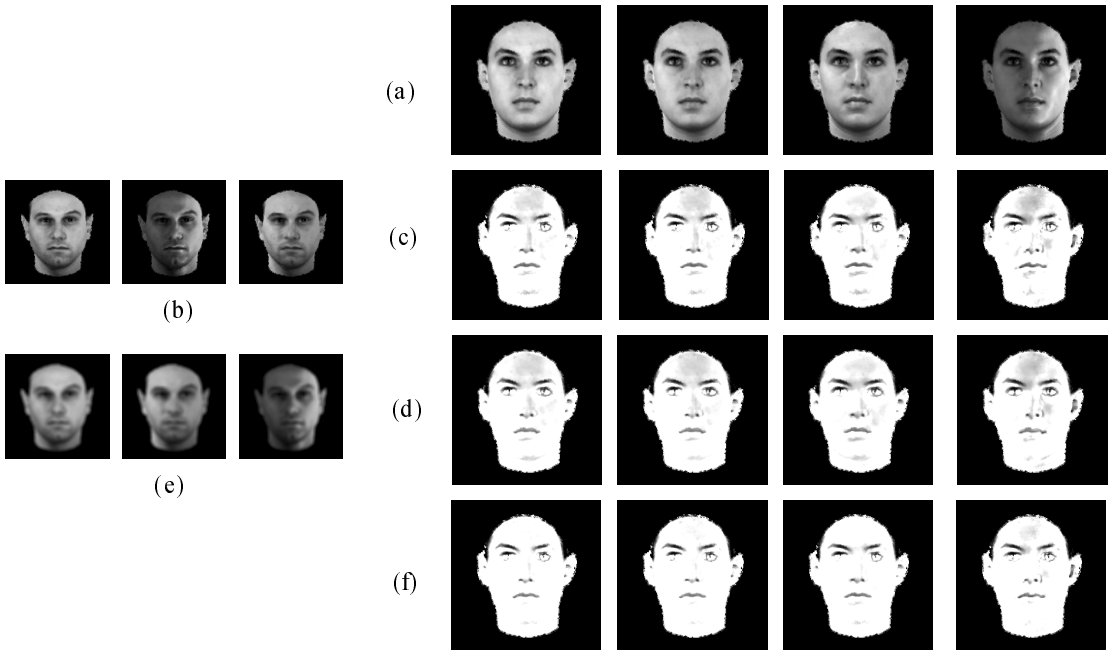


Figure 6.4. Testing accuracy of Theorem 1 against the size of the bootstrap set. (a) Original images taken under 4 distinct light conditions. (b) Bootstrap set of $N = 1$ objects used for generating the Q-images of (a) displayed in row (c). Note that the quotient images are not strictly invariant as they change with the illumination. (d) Q-images of the bootstrap set ($N = 1$) displayed in (e). Note that the bootstrap set is blurred in order to test whether using the "average" object when $N > 1$ makes a difference compared to the machinery described in Theorem 1. We see that blurred images do not improve the invariance of the Q-images. (f) Q-images of (a) against the object (b) but where the coefficient vector x was recovered using the $N = 10$ bootstrap set of Fig. 6.1. The comparison should be made between rows (c) and (f). Note that in (f) the images are invariant to changing illumination more so than in (c).

Figure 6.5. Q-images should be invariant to the 3 illumination conditions of the database images, as long as they span a 3 Dimensional subspace. The 3 Q-images were generated against different bootstrap sets of the same 10 objects but of different triplets of light sources. Note that the novel object is not part of the original database of 200 objects, but of a member of our lab.

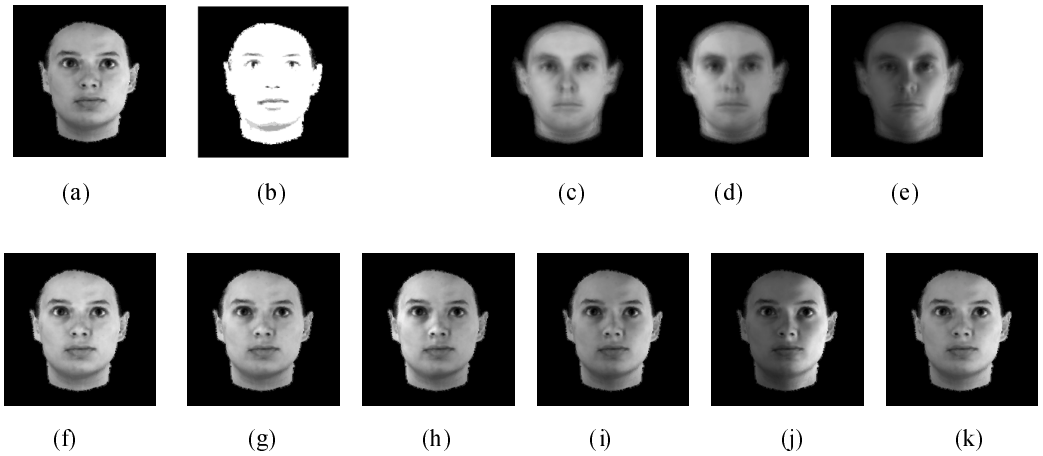


Figure 6.6. Image Synthesis Example. (a) Original image and its quotient image (b) from the $N = 10$ bootstrap set. The quotient image is generated relative to the average object of the bootstrap set shown in (c),(d) and (e). Images (f) through (k) are synthetic images created from (b) and (c),(d), (e) using Proposition 1.

In Figs. 6.6 and 6.7 we demonstrate the results of image synthesis from a single input image and the bootstrap set. Note the quality and the comparison between results of bootstrap size $N = 10$ and $N = 2$ (there are differences but relatively small).

To complete this set of experiments, we have conducted the following tests, which measure almost directly the accuracy of the light coefficients. Given 3 images of the same subject, y_1, y_2, y_3 under 3 distinct light coefficients (base images) and a fourth image of the subject taken under a different light y_s, we term "real" the coefficients needed to generate the fourth image out of the first three. That is $y_s = \sum_{i=1}^{3} \beta_i y_i$ where $\beta_1, \beta_2, \beta_3$ are the "real" light coefficients. The Belhumeur database which was taken under controlled lighting conditions, (precisely measured) was used here, to display a comparison between the "real" light coefficients (the β-s) to those that were calculated using Q-image algorithm (noted here by x_1, x_2, x_3). The comparison can be seen graphically in Figure 6.8. For clarity, each calculated (Q-image) coefficient was compared separately to the "real" one. The comparison that was done for a set of varied illuminations, detailed in Table 6.1, shows that though the calculated coefficients and the "real" do not actually overlap, the differences are comparatively small. The variations between

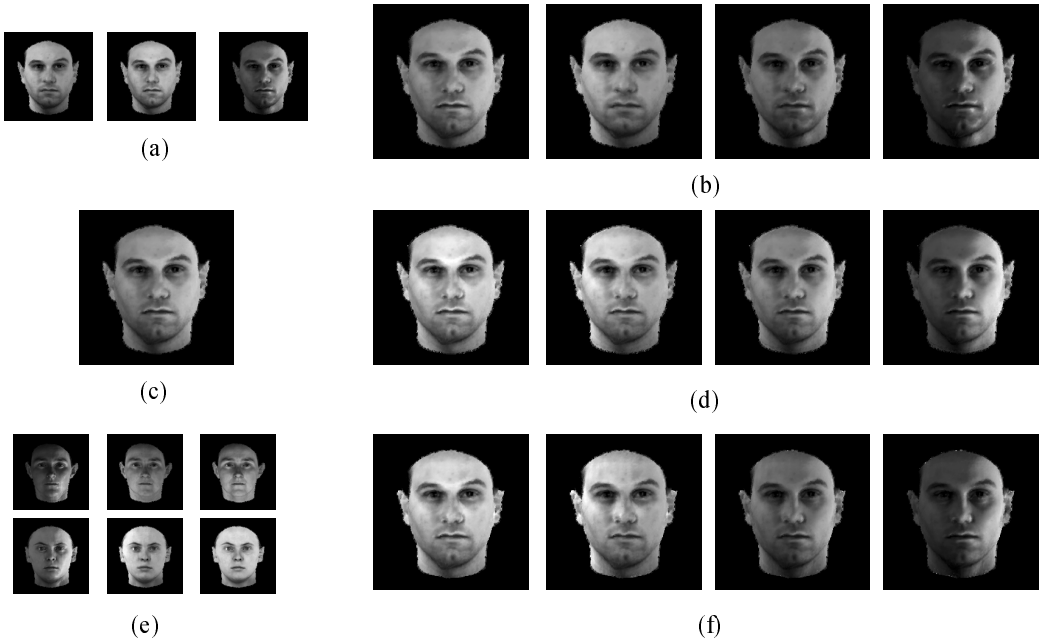


Figure 6.7. Image synthesis examples. (a) Original images under 3 distinct lighting conditions and the synthesized images (b) using linear combinations of those 3 images. The synthesized images using the original single image (c) and a $N = 10$ bootstrap set are shown in (d). Finally, (e) is an $N = 2$ bootstrap set for generating the synthesized images (f) from the single original image (c).

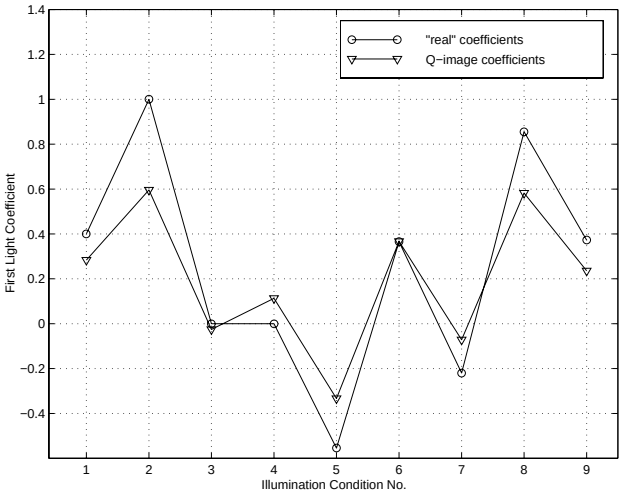
the coefficients are presumably due to shape variations between the faces, which were not taken into consideration in the algorithm. Varying the bootstrap set size in this experiment did not effect the coefficient accuracy significantly. It is presumably due to large variations between the bootstrap set objects compared to its size.

No.	Horizontal Angle	Vertical Angle
1	0	0
2	0	-35
3	-50	0
4	50	0
5	0	45
6	0	10
7	0	20
8	0	-20
9	-10	0

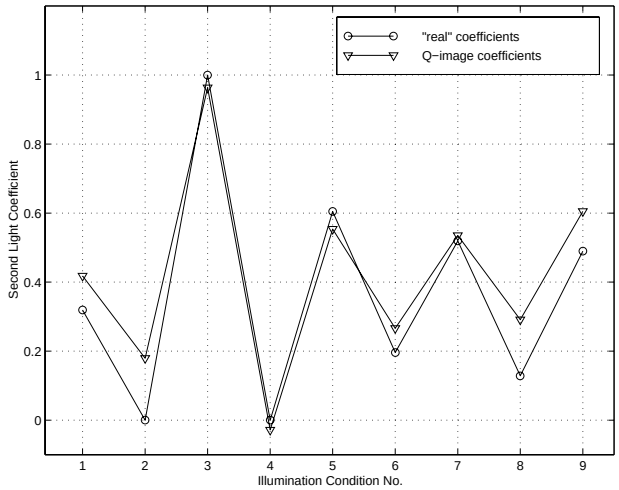
Table 6.1. Illumination Directions.

6.1.6 Rank of Estimation Matrix

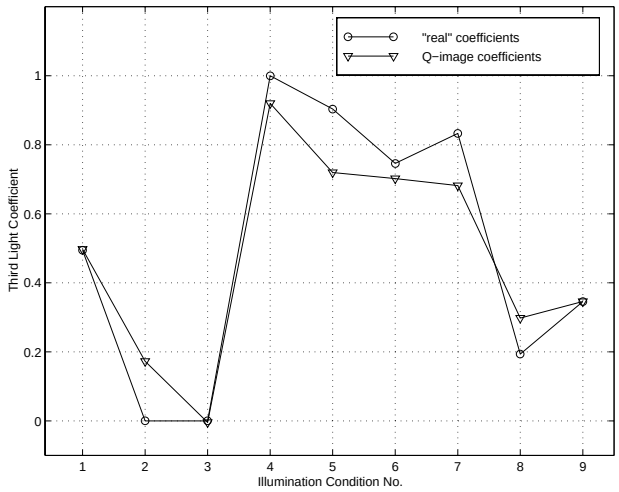
In the previous subsection we have tested the accuracy of the light coefficients x_1, x_2, x_3. Their correctness depends on exact estimation of the scale coefficients $\alpha_1 \ldots \alpha_N$. The α-s can be derived quite easily from a set of linear equations, termed an *Estimation Matrix*, (see eqn 5.2, in Chapter 5). This estimation matrix is noted by F. To solve for the α-s one can use SVD. We have found out,



(a)



(b)



(c)

Figure 6.8. Comparison between the light coefficients for 9 different illumination conditions, detailed in Table 6.1, using images from Belhumeur database (see Fig. 6.2). For clarity, the comparison is done for each of the 3 light coefficients saperately.

empirically, for all the databases sets we have worked with that the first $N-1$ singular values are almost equal, while the last one is considerably smaller, regardless of the size of N. This suggests that the rank of F is $N-1$, that is we have one degree of freedom in determining the scale of the α-s. Since the accuracy of x is conditioned by a correct scaling (see Theorem 2), it should be done wisely. We have chosen to scale the α-s so that $\sum_{i=1}^{N} \alpha_i = N$. This is a point of fragility in the algorithm: for example, in case of only one object in the bootstrap set, there is no scale compensation for the albedo, so we actually assume that the albedos of the novel object ρ_y and of the bootstrap set object ρ_a are similar. However, in the common cases this is the best estimation that can be made.
The next subsection relates to the meaning of the α-s.

6.1.7 The Meaning of the Scale Coefficients

In Chapter 5 we have stated (by definition) that the albedo ρ_y of object y is rationally spanned by the bootstrap set if there exist coefficients $\alpha_1 \dots \alpha_N$ such that

$$\rho_y = \frac{\rho_1^2 \dots \rho_N^2}{\alpha_1 \rho_1 \dots \alpha_N \rho_N}$$

We would now like to ask in what way these α-s function as scale coefficients. The expression above suggests that the α-s are correlated with the average intensity, since α_i is a scalar that multiplies the reflectance values $\rho_i(x,y)$ of each point (x,y) of the i-th bootstrap set of object images. To check that we have averaged the pixel intensities of each image. Since each bootstrap set object is represented by 3 images, we than take the mean of the 3 averaged values and plot it against the corresponding α as shown in Figure 6.10 according the values in Table 6.2. The graph is almost a straight line, i.e. there is an indubitable correlation between the α-s and the average intensity of the corresponding object images, checked for a bootstrap set of 9 objects. This supports (empirically) our assumption. In Figure 6.9 the bootstrap set images are displayed with the corresponding α-s. Note, that the darkest images, have the lowest α-s. Yet, if the bootstrap set images have similar total intensities and the subjects do not differ in their skin color, the values of the α are with small variance from one. That is, computation time can be saved setting all the α-s to one in advance. Smarter option would be to calculate the average intensities of the bootstrap images and to use it to determine the α-s in advance, scaling so that $\sum_{i=1}^{N} \alpha_i = N$.

average intensity	α
71	0.76
86	0.93
88	0.95
89	0.96
94	1.02
95	1.04
98	1.05
107	1.15
107	1.15

Table 6.2. α versus avarge image intensitiy.

6.2 Synthesis Results

We have used the Belhumer database, that was taken under supervised illumination conditions (that were carefully measured and reported) to test the similarities between original images and synthesized

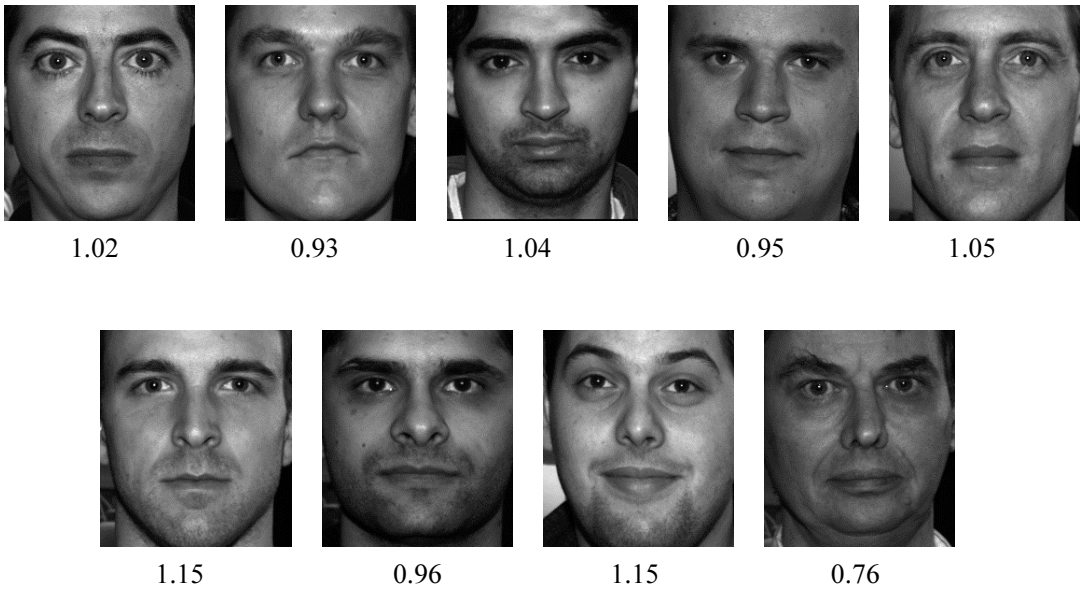


Figure 6.9. Selection of bootstrap set images from Belhumeur database. The value of α is written under each image. Notice that the darkest images have the lowest α-s.

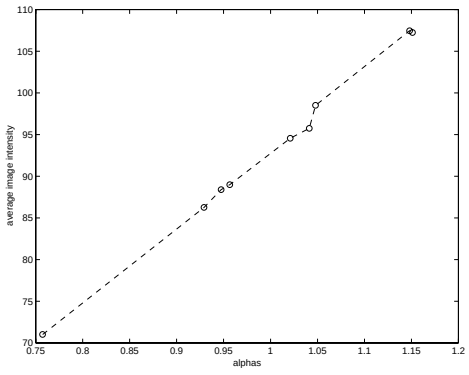


Figure 6.10. The value of α is plotted against the averaged images intensities of each of the corresponding bootstrap set object (taken from belhumer database) The almost straight line indicates high correlation

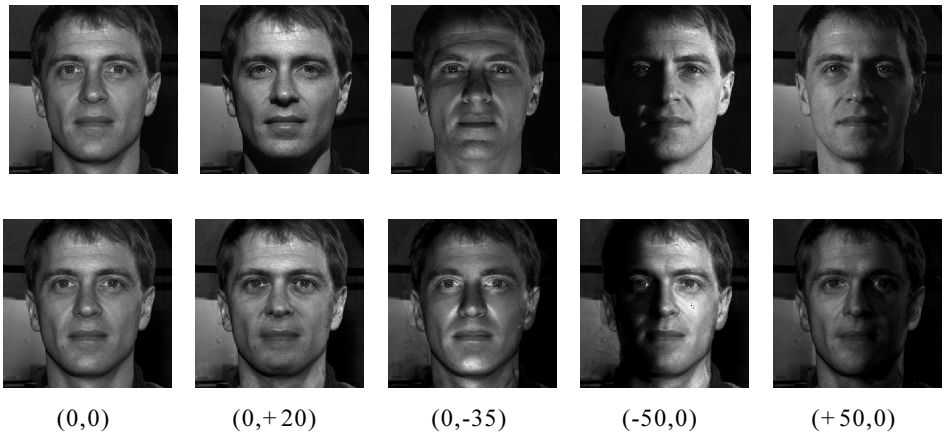


Figure 6.11. The five images in the first row are original images taken from Belhumeur database illuminated from the center, up, down ,right and left correspondingly. The exact values of the light directions are written in the third row in correspondance (The left values in the brackets stand for the horizontal angle, and the right ones for the vertical). The synthesized images in the third row simulate these illumination conditions.

ones. Five images of one of the subject, $y_1 \dots y_5$, were selected from the database, in which the illumination directions are center, up, down, right and left, as can be seen in the first row of Figure 6.11. The exact values of the light directions are displayed in the third row in correspondence. Now, taking 3 images of this subject with distinct illuminations, we've computed the light coefficients for each of the 5 original images. Using these calculated light coefficients we have synthesized new images of subject a given only one image of a and a bootstrap set of 4 other objects taken under the same 3 base illuminations. The resulting synthesized images are displayed in the second row of Figure 6.11. The bootstrap set is shown in Figure 6.2.

Synthesizing images for practical uses, one can always supervise the quality of the images in the bootstrap set. Yet, it is still challenging to test the algorithm with low quality images. The next experiment was imperimented with another bootstrap set shown in Fig. 6.12a. A bootstrap set of three objects varying in hair-style, uncropped, and generally taken under much less attention compared to the bootstrap set of Fig. 6.1 or Fig. 6.2 is sufficient, nevertheless, to generate quite reasonable re-renderings as shown in Fig. 6.12d. The degradation is indeed graceful and affects mainly the degree of illumination changes, not as much the quality of the resulting image (compared to the source image).

So far we have imperimented with objects and their images from the same database (Vetter's database, Belhuemer's database). Even though the input image is of an object outside the bootstrap set, there is still an advantage in having all the images taken with the same camera, same conditions and same quality level. Our next imperiments were designed to test the algorithm on images taken from sporadic sources, such as magazines or the Web. The bootstrap set in all imperiments is the one displayed in Fig. 6.1.

Fig. 6.13 shows four novel (color) images of celebrities (from magazines) and the result of the synthesis procedure. These images are clearly outside the circle of images of the original database of Vetter, for example the images are not cropped for hair adjustment and the facial details are visibly different from those in the bootstrap set.

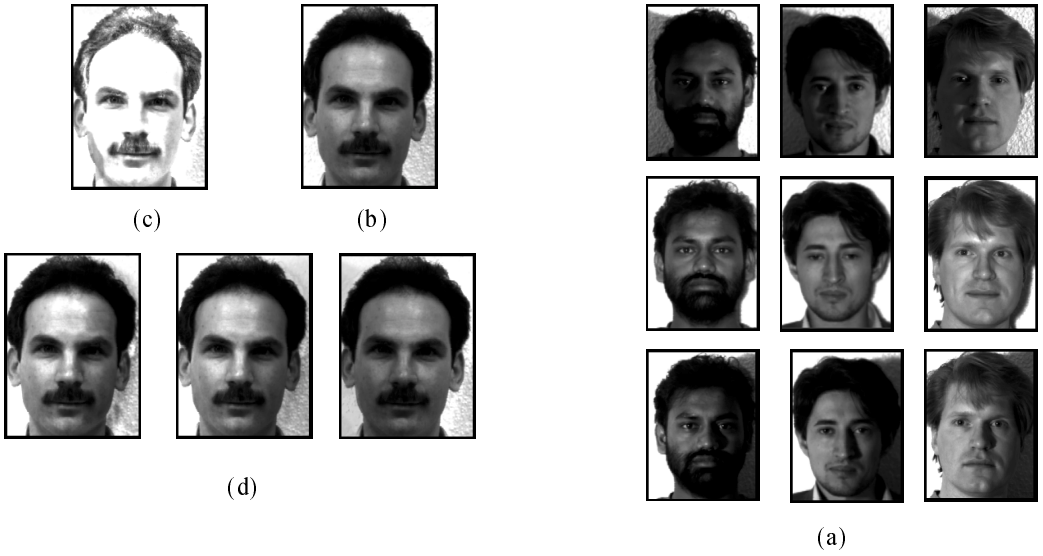


Figure 6.12. Image synthesis using other, lower quality, bootstrap sets (Yale data sets). The bootstrap set ($N = 3$) is shown in (a). Note that the objects vary considerably in appearance (hair style and facial hair) and are thus less controlled as in Vetter's data set. The source image (b), its quotient image (c) and synthesized images (d).

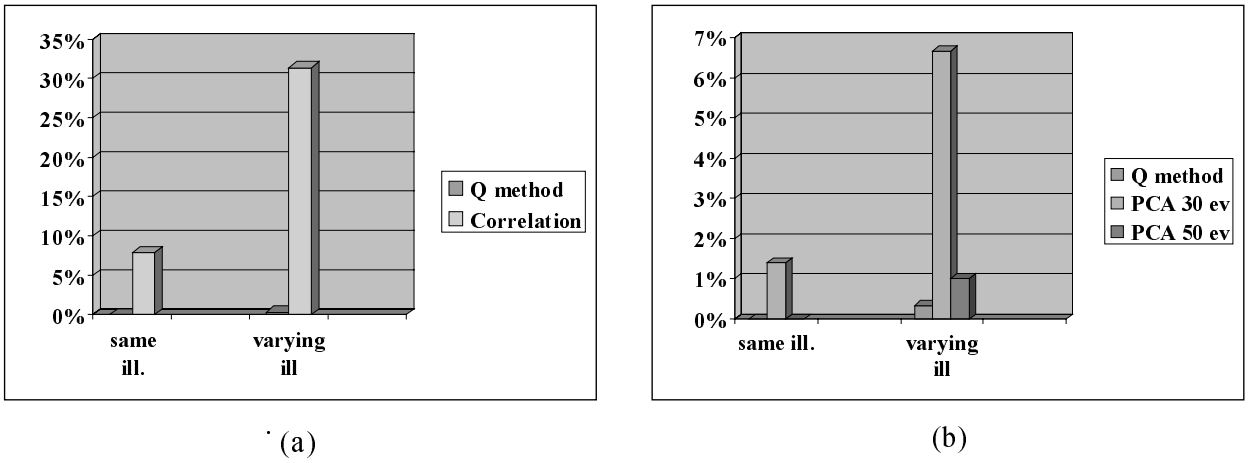
6.3 Recognition Results

The Q-images are illumination invariant signatures of the objects in the class. We can therefore make use of the invariance property for purposes of recognition. Vetter's data base contains 200 faces each under 9 lighting conditions, making a total of 1800 images. We used a bootstrap set of 20 objects (60 images) and created the Q-images of all the 200 objects — these 200 images serve as the database, we refer to as Q-database, for purposes of recognition. Given any of the 1800 source images, its Q-image is created from the bootstrap set and matched (by correlation) against the Q-database while searching for the best match.

We made two tests (summarized in Fig. 6.14). In the first test the Q-database was generated from images under the same illumination (we have 9 images per object in Vetter's database). The results of recognition were compared to correlation were the database for correlation where those images used for creating the Q-database. The match against the Q-database was error free (0%). The match against the original images, instead of the Q-images, had 142 mismatches (7.8%). In the second test the images used for creating the Q-database were drawn randomly from the set of 9 images (per object). The match against the Q-database produced only 6 mismatches (0.33%), whereas the match against the original images produced 565 mismatches (31.39%). The sharp increase in the rate of mismatches for the regular correlation approach is due to the dominance of illumination effects on the overall brightness distribution of the image (cf. [Shashua, 1997, Adini et al., 1997]).

We also made a comparison against the "eigenfaces" approach explained in **??**, which involves representing the database by its Principle Components (PCA). In the first test, PCA was applied to the bootstrap set (60 images) and 180 additional images, one per object. In the first test the additional images were all under the same illumination, and in the second test they were drawn randomly from the set of 9 images per object. The recognition performance depends on the number of principle components. With 30 principle components (out of 240) the first test had 25 mismatches (1.4%), and the second test had 120 mismatches (6.6%). The performance peaks around 50 principle components in which case the first test was error free (like in the Q-image method), and the second test had 18 mismatches (1%).

(a) (b) (c) (d) (e) (f) (g)



(a) (b)

Figure 6.14. Recognition results on Vetter's database of 1800 face images. We compare the Q-image method with correlation and Eigenfaces. See text for details.

To summarize, in all recognition tests, except one test of equal performance with PCA, the Q-image outperforms and in some cases in a significant manner, conventional class-based approaches.

6.4 Other Routes for a Signature Image?

The quotient image approach is based on the idea that an illumination invariant image $Q = \rho_y/\rho_a$ can be used to map the image space of object $\mathbf{a}$ to the image space of object $\mathbf{y}$ using a single image y_s of $\mathbf{y}$. The equation $(\sum_j x_j a_j) \otimes Q$ generates the image space of $\mathbf{y}$ (Proposition 1). There are two points worth making.

First, Q is analogous to an "error correction term". However, it is important to distinguish between error correction and an illumination invariant term. For example, let $\hat{y}$ be the reconstructed image of y_s from the bootstrap set (after solving for x, α_i that minimizes eqn. 4.1 in the "reconstructionist" approach), and let $\bar{Q}$ be defined so that $y_s = \hat{y} \otimes \bar{Q}$. *There is no reason to expect that $\bar{Q}$ would be illumination invariant.* This is demonstrated in Fig. 6.15b showing that the $\bar{Q}$ images are not invariant to changing illumination. In other words, we would not obtain an admissible image space of $\mathbf{y}$, or correct re-rendering, if we simply correct for the reconstruction error by a Cartesian product with $\bar{Q}$.

Second, notice that the optimization criteria described in Theorem 1 involves a somewhat complex definition of what constitutes a "family" of albedo functions (rational span). This is unlike the more intuitive definition, that would typically adopt under such circumstances, that albedo functions are closed under linear combinations (the definition adopted in the optimization criteria behind eqn. 4.1 for the "reconstructionist" approach). However, the rational span definition has an important role because through it we were able to remove the intrinsic bilinearity among the illumination parameters $x = (x_1, x_2, x_3)$ and the albedo parameters $\alpha_1, ..., \alpha_N$ and obtain a linear system for $N+3$ variables (instead of $3N$ if the linear span definition were to be adopted). The importance of all this depends on the numerical behavior of the system. In principle, however, one could solve for x from eqn. 4.1 and use it for obtaining the quotient image as defined in Proposition 1. In other words, in the algorithm described in the previous section, simply replace steps 2–4 with the procedure described in Chapter 4 for obtaining x. We expect a degradation in performance due to numerical considerations (due to the enlargement of parameter space). The results of doing so are illustrated in Fig. 6.15c. The quotient images clearly show a dependence on illumination change, indicating that the parameters x_1, x_2, x_3 were not recovered well.

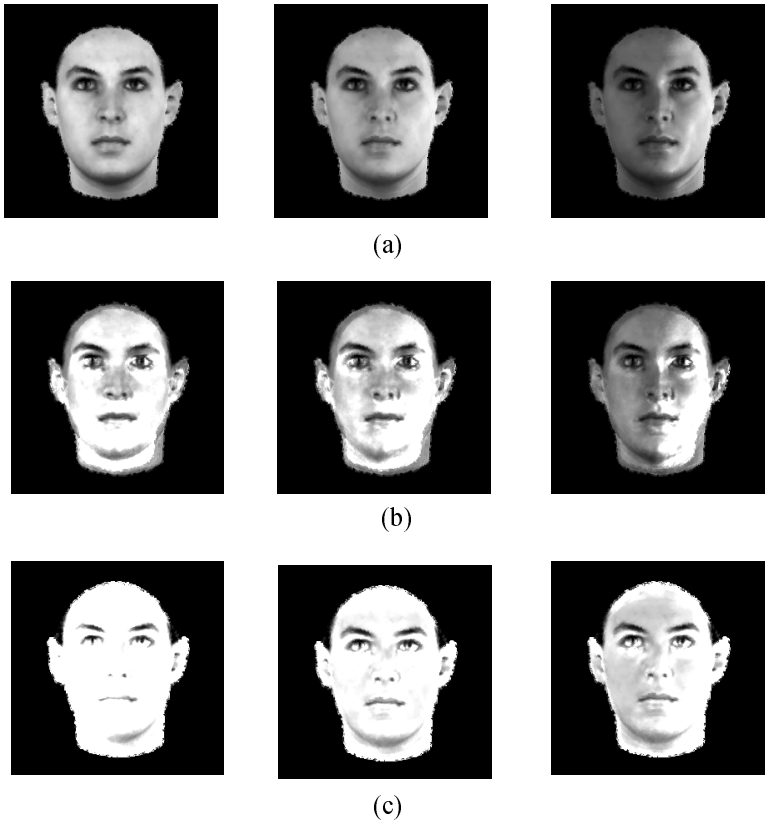


Figure 6.15. Alternatives approaches for a quotient image. (a) original images under varying illumination. (b) Quotient images defined as a multiplicative "error" image, i.e., the ratio of the original image and the least-squares reconstructed image from the bootstrap set. Note that the resulting quotient images are not illumination invariant. (c) Quotient images defined by Proposition 1 where x is the minima of eqn. 4.1 (instead of eqn. 5.1). Again the images are not illumination invariant.

In summary, the combination of an illumination invariant correction term (the quotient image) and a simple optimization criteria (eqn. 1) — with the price of somewhat complicating the definition of when albedos form a "family" — gives rise to both practical and a proven correct procedure for class-based re-rendering (under the terms stated of ideal class definition and Lambertian surfaces).

Chapter 7

Summary and Future Research Directions

We have presented a class-based, image-based, re-rendering and recognition method. The key element of our approach was to show that under fairly general circumstances it is possible to extract from a small set of example images an illumination invariant "signature" image per novel object of the class from a single input image alone. We have proven our results (under the "imaginary" world of ideal class assumption) and demonstrated the applicability of our algorithm on the class of real pictures of human faces. In other words, we have shown that in practice a remarkably small number of sample images of human frontal faces (in some of our experiments images of two objects were sufficient for making a database) can generate photo-realistic re-rendering of new objects from single images.
We have also shown that the synthesis algorithm can be applicable for images taken from different sources, with different qualities. Color images can also be "re-lighted" and look natural even by using a gary-level bootstrap set.
In spite of the fact that, in comparison of the synthesized images with the real ones, one can detect the differences, i.e. the illumination conditions cannot be imitated precisely, the perception of the artificially generated images is as if they were real. Moreover, the preciseness in determination of the light coefficients can be drastically improved increasing the size of the bootstrap set. In addition, small variations of the light coefficients, supervised manually should yield the desirable results. Finally, one should always keep in mind that non-linear light effects such as highlights or cast shadows cannot be simulated, and an expansion of the algorithm is needed to handle these phenomena.
The "signature" image presents an alternative representation of an object due to its invariance to illumination variations, while, yet preserving the object unique features. The recognition algorithm offered in this thesis, was tested on a database of 1800 images of 200 subjects, taken under 9 different lights. The subjects were of the same race and age, wore tight sweeming caps, had no glasses, makeup of facial hair and all the images were taken with a black background. Thus the only cues for recognition were the face features. The results outperform by far conventional methods, as is indicated by the very low error rate (below 0.5%). The algorithm applicability is also expressed in terms of simplicity, low complexity and short computational time. Furthermore the demand for precise determination of the light coefficients, which is essential in synthesis tasks, can be weakened.
The ideas presented in this paper can, without too much difficulty, be turned into a system for image composing and re-lighting of general faces, with very high quality of performance. To that end, further implementation elements may be required, such as using collections of bootstrap sets (while choosing among them manually or automatically using sparse optimization approaches like Support Vector Machines [Vapnik, 1995]), and automatic or semi-automatic tools for morphing the bootstrap set onto the novel image in order to better compensate for changes of shape (such as [Vetter et al., 1997]).
The next challenging step might be integration between three sources of variations: Reflectance prop-

erties, using objects of the class, illumination conditions and geometric variations. That is to uniquely define an object of a class disregarding geometric or photometric variations, or to simulate variety of viewing conditions given only one image of an object, based on the appearance of other objects of the class.

[Adini et al., 1997] Adini, Y., Moses, Y., and Ullman, S. (1997). Face recognition: the problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732.

[Ahmed and Goldstein, 1975] Ahmed, N. and Goldstein, M. (1975). *Orthogonal Transform for Digital Signal Processing*. Springer-Verlag, New York.

[Ash and Gardner, 1975] Ash, R. and Gardner, M. (1975). *Topics in Stochastic Processes*. Academic, New York.

[Atick et al., 1997] Atick, J., Griffin, P., and Redlich, N. (1997). Statistical approach to shape-from-shading: deriving 3d face surfaces from single 2d images. *Neural Computation*.

[Basri, 1996] Basri, R. (1996). Recognition by prototypes. *International Journal of Computer Vision*, 19(2):147–168.

[Belhumer and Kriegman, 1997] Belhumer, P. and Kriegman, D. (1997). What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, pages 1–16.

[Belhumeur et al., 1996] Belhumeur, P., Hespanha, J., and Kriegman, D. (1996). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In *Proceedings of the European Conference on Computer Vision*, pages 45–58.

[Belhumeur et al., 1997] Belhumeur, P., Kriegman, D., and Yuille, A. (1997). The bas-relief ambiguity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1040–1046.

[Beymer and Poggio, 1996] Beymer, D. and Poggio, T. (1996). Image representations for visual learning. *Science*, 272:1905–1909.

[Chellapa et al., 1995] Chellapa, R., Wilson, C., and Sirohey, S. (1995). Human and machine recognition of faces: A survey. In *Proceeding of the IEEE*, volume 83, pages 704–740.

[Coren and Ward, 1989] Coren, S. and Ward, L. (1989). *The Constancies, In Sensation and Perception*, chapter 14, pages 403–425. 3rd edition.

[Devijver and Kittler, 1982] Devijver, P. and Kittler, J. (1982). *Pattern Recognition: A statistical approach*, chapter 9. Prentice Hall International, London.

[Edelman, 1995] Edelman, S. (1995). Class similarity and viewpoint invariance in the recognition of 3d objects. *Biological Cybernetics*, 72:207–220.

[Fan and Wolff, 1997] Fan, J. and Wolff, L. (1997). Surface curvature and shape reconstruction from unknown multiple illumination and integrability. *Computer Vision and Image Understanding*, 65(2):347–359.

[Farah, 1988] Farah, M. (1988). Is visual imagery really visual? overlooked evidence from neuropsychology. *Psychological Review*, 95(3):307–317.

[Field, 1994] Field, D. (1994). What is the goal of sensory coding? *Neural Computation*, 6(4):559–601.

[Foldiak, 1990] Foldiak, P. (1990). Forming sparse representation by local anti-hebbian learning. *Biological Cybernetics*, 64:165–170.

[Freeman and Tenenbaum, 1997] Freeman, W. and Tenenbaum, J. (1997). Learning bilinear models for two-factor problems in vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 554–560.

[Fukunaga, 1989] Fukunaga, K. (1989). *Introduction to Statistical Pattern Recognition*. New York: academic.

[Georghiades et al., 1998] Georghiades, A., Kriegman, D., and Belhumeur, P. (1998). Illumination cones for recognition under variable lighting:faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 52–59.

[Hallinan, 1994] Hallinan, P. (1994). A low-dimentional representation of human faces for arbitrary lightening conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–999.

[Horn, 1977] Horn, B. (1977). Image intensity understanding. *Artificial Intelligence*, 8:201–231.

[Horn, 1986] Horn, B. (1986). *Robot Vision*. MIT Press, Cambridge, Mass.

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

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

[Kohonen, 1988] Kohonen, T. (1988). *Self-Orginizing and Associative Momery*. Springer-Verlag, Berlin.

[Meer and Leedan, 1998] Meer, P. and Leedan, Y. (1998). Estimation with bilinear constraints in computer vision. In *Proceedings of the International Conference on Computer Vision*, pages 733–738, Bombay, India.

[Meunier and Nadal, 1995] Meunier, C. and Nadal, J. (1995). Sparsely coded neural networks. In Arbib, M., editor, *The Handbook of Brain Theory and Neural Networks*, pages 899–901. MIT Press, Cambridge, Massachusetts.

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

[Moses et al., 1996] Moses, Y., Ullman, S., and Edelman, S. (1996). Generalization to novel images in upright and inverted faces. *Perception*, 25:443–461.

[Nalwa, 1993] Nalwa, V. (1993). *A Guided Tour of Computer Vision*. Addison-Weseley Publishing Company.

[Nayar and Bolle, 1995] Nayar, S. and Bolle, R. (1995). Reflectance based object recognition. *International Journal of Computer Vision*.

[Nayar et al., 1991] Nayar, S., Ikeuchi, K., and Kanade, T. (1991). Surface reflection: Physical and geometrical perspective. *pami*, 13(7):611–634.

[Nimeroff et al., 1994] Nimeroff, J., Simoncelli, E., and Dorsey, J. (1994). Efficient re-rendering of naturally illuminated environments. In *Proceedings of the Fifth Annual Eurographics Symposium on Rendering*, Darmstadt Germany.

[Oja, 1992] Oja, E. (1992). Principle components, minor components and linear neural networks. *Neural Networks*, 5:927–935.

[Oja, 1995] Oja, E. (1995). Principle component analysis. In Arbib, M., editor, *The Handbook of Brain Theory and Neural Networks*, pages 753–756. MIT Press, Cambridge, Massachusetts.

[Olshausen and Field, 1995] Olshausen, B. and Field, D. (1995). Sparse coding of natural images produces localized, oriented, bandpass receptive fields. Technical Report CCN-110-95, Department of Psychology, Cornell University, Itaca, New York.

[Pentland, 1982] Pentland, A. (1982). Finding the illuminant direction. *Journal of the Optical Society of America*, 72:448–455.

[Pentland, 1984] Pentland, A. (1984). Local shading analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:170–187.

[Poggio and Vetter, 1992] Poggio, T. and Vetter, T. (1992). Recognition and structure from one 2d model view: Observations on prototypes, object classes and symmetries. Technical Report AI Memo 1347, MIT.

[Pontil and Verri,] Pontil, M. and Verri, A. (?). 3-d object recognition with support vector machnies. *?*

[Rittenhouse, 1786] Rittenhouse, D. (1786). Explenation of an optical deception. *Transaction of the American Philosophy Society*, 2:37–42.

[Rosch, 1973] Rosch, E. (1973). *On the Internal Structure of Perceptual and Semantic Categories*. ?

[Rosch and Mervis, 1975] Rosch, E. and Mervis, C. (1975). Familiy resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, pages 573–650.

[Rosch et al., 1976] Rosch, E., Mervis, C., W.D. Gray, D. J., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

[Rowland and Perrett, 1995] Rowland, D. and Perrett, D. (1995). Manipulating facial appearance through shape and color. *IEEE Computer Graphics and Applications*, pages 70–76.

[Sali and Ullman, 1998] Sali, E. and Ullman, S. (1998). Recognizing novel 3-d objects under new illumination and viewing position using a small number of examples. In *Proceedings of the International Conference on Computer Vision*, pages 153–161.

[Schoeneman et al., 1993] Schoeneman, C., Dorsey, J., Smits, B., Arvo, J., and Greenberg, D. (1993). Painting with light. In *Computer Graphics Proceedings, Annual Conference Series*, pages 143–146.

[Shashua, 1992] Shashua, A. (1992). Illumination and view position in 3D visual recognition. In J.E. Moody, S. H. and Lippmann, R., editors, *Advances in Neural Information Processing Systems 4*, pages 404–411. San Mateo, CA: Morgan Kaufmann Publishers. Proceedings of the fourth annual conference NIPS, Dec. 1991, Denver, CO.

[Shashua, 1995] Shashua, A. (1995). Algebric functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789.

[Shashua, 1997] Shashua, A. (1997). On photometric issues in 3D visual recognition from a single 2D image. *International Journal of Computer Vision*, 21:99–122.

[Shepard and Cooper, 1982] Shepard, R. and Cooper, L. (1982). *Mental Images and their Transformation*. Cambridge MA: MIT Press.

[Sirovich and Kirby, 1987] Sirovich, L. and Kirby, M. (1987). Low dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524.

[Smith, 1978] Smith, C. (1978). Color gamut transformation pairs. *Computer Graphics*, 12:12–19.

[Tarr and Gauthier, 1998] Tarr, M. J. and Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition*, 67(1-2):71–108.

[Tenenbaum and Freeman, 1997] Tenenbaum, J. and Freeman, W. (1997). Saparating style and content. In M. Mozer, M. J. and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 662–668.

[Tomasi and Kanade, 1992] Tomasi, C. and Kanade, T. (1992). Shape and motion from image streams – a factorization method. *International Journal of Computer Vision*, 9(2):137–154.

[Turk and Pentland, 1991] Turk, M. and Pentland, A. (1991). Eigen faces for recognition. *J. of Cognitive Neuroscience*, 3(1).

[Ulman and Basri, 1991] Ulman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006.

[Vapnik, 1995] Vapnik, V. (1995). *The nature of statistical learning*. Springer.

[Vetter and Blanz, 1998] Vetter, T. and Blanz, V. (1998). Estimating coloured 3d face models from single images: an example based approach. In *Proceedings of the European Conference on Computer Vision*, pages 499–513.

[Vetter et al., 1997] Vetter, T., Jones, M., and Poggio, T. (1997). A bootstrapping algorithm for learning linear models of object classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 40–46.

[Vetter and Poggio, 1996] Vetter, T. and Poggio, T. (1996). Image synthesis from a single example view. In *Proceedings of the European Conference on Computer Vision*.

[Vetter and Poggio, 1997] Vetter, T. and Poggio, T. (1997). Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):733–742.

[Wolff and Angelopoulou, 1994] Wolff, L. and Angelopoulou, E. (1994). 3-d stereo using photometric ratios. In *Proceedings of the European Conference on Computer Vision*, volume 801, pages 247–258.

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

[Yuille and Snow, 1997] Yuille, A. and Snow, D. (1997). Shape and albedo from multiple images using integrability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 158–164.