Stony Brook University

Illumination and Geometry Inference Using Graphical Models

A Dissertation Presented
by
Alexandros Panagopoulos

to
The Graduate School
in Partial Fulfillment of the
Requirements
for the Degree of

Doctor of Philosophy
in
Computer Science

Stony Brook University

December 2011

Stony Brook University
The Graduate School

Alexandros Panagopoulos

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Dimitris Samaras - Dissertation Advisor
Associate Professor, Computer Science Department

Tamara Berg - Chairperson of Defense
Assistant Professor, Computer Science Department

Alexander Berg
Assistant Professor, Computer Science Department

David Forsyth
Professor, Computer Science Department, University of Illinois at Urbana-Champaign

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Illumination and Geometry Inference Using Graphical Models

by

Alexandros Panagopoulos

Doctor of Philosophy

in

Computer Science

Stony Brook University

2011

Image formation is a function of three components: scene geometry, surface reflectance and illumination. Estimation of one or more of these components from an image gives rise to inverse rendering problems, such as shape reconstruction or illumination estimation, which are the two major problems of interest in this thesis. We formulate such problems in a way that attempts to bridge the gap between low-level approaches based on the physical laws governing image formation and higher-level models that examine images in a statistical way. We take advantage of the powerful formalism offered by graphical models, which lead to modular frameworks and offer powerful discrete optimization techniques. We first focus on the problem of illumination estimation from a single image, utilizing the information in cast shadows. We start by describing a method to extract cast shadows from an image. We then present three approaches to illumination estimation from shadows: The first models illumination as a mixture of distributions to robustly estimate illumination. The second associates illumination not with pixel intensities but with the existence of shadow edges. The third approach unifies the previous ideas in a Markov Random Field (MRF) framework. Such a model is robust to coarse or incomplete knowledge of geometry, while it can also incorporate geometric parameters, allowing us to jointly infer three major components of the problem: the cast shadows, illumination and geometry. Geometry inference from the information contained in cast shadows can only be coarse, however. We subsequently focus on the problem of inferring geometry from the shading variations in an image. We take a data-driven approach, constructing a dictionary of geometric primitives. To reconstruct an image, we combine local hypotheses from this dictionary in an MRF model. We demonstrate that this approach can effectively reconstruct 3D shapes from real photographs, while removing several important assumptions of previous approaches.

Contents

List of Figures

List of Tables

Acknowledgments

I would first like to thank my advisor, Dimitris Samaras, for his help, support and guidance, without which the work in this thesis would not have been possible. During the years i worked with him, he taught me how to identify interesting research problems, how to approach them in search for solutions, but also how to persist and handle the challenges of research and life in general. I would also like to thank the members of my dissertation committee, Prof. Tamara Berg, Prof. Alexander Berg and Prof. David Forsyth, who provided me with excellent feedback for my future research. Another person without whom this dissertation would not have been possible is my girlfriend Jin-kang, who warmly and patiently supported and tolerated me for the last four years. Her support has been a valuable source of strength for me.

The friends I made at Stony Brook were certainly another force that helped me substantially to cope with the challenges I faced over the years, while working for this dissertation. I would also like to thank my labmates, who provided an environment for both intellectual stimulation and friendly conversation. Last but not least, my parents, who set the wheels in motion a long time ago and offered me help and support since.

Publications

Publications from this thesis:

- (under submission) "Simultaneous Cast Shadows, Illumination & Geometry Inference Using Hypergraphs", A. Panagopoulos, C. Wang, D. Samaras, N. Paragios, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence
- "Illumination Estimation from Shadow Borders", Alexandros Panagopoulos, Tomas Yago Vicente, Dimitris Samaras, 3rd IEEE Workshop on Color and Photometry in Computer Vision (in conjunction with ICCV'11), 2011
- A.Panagopoulos, C.Wang, D.Samaras, N.Paragios, "Illumination Estimation and Cast Shadow Detection through a Higher-order Graphical Model", in CVPR 2011
- A.Panagopoulos, C.Wang, D.Samaras, N.Paragios, "Estimating Shadows with the Bright Channel Cue", in CRICV2010 (in conjunction with ECCV'10)
- A.Panagopoulos, D.Samaras, N.Paragios, "Robust Shadow and Illumination Estimation Using a Mixture Model", in CVPR 2009

Chapter 1

Introduction

The computer vision problem can be examined at different conceptual levels. At a lower level, one could try to model the image formation process based on the laws of physics that govern it. At a higher level, one could attempt to identify objects or extract scene properties based on statistical models of the appearance of objects in the image, disregarding many components of the process that leads to the specific appearance of the object in the image. It is clear that none of these approaches is a complete treatment of the problem: trying to extract information from an image by modeling the physical laws that govern its formation necessitates many restricting assumptions and accurate knowledge of components of the image formation other than the input image. On the other hand, higher-level computer vision tasks treat many of the components of the image formation process, such as illumination, as noise that has to be ignored. The result is often compromised performance.

In this thesis we are interested in the image formation process. This process is the interaction of geometry, reflectance and illumination that leads to the formation of an image. Attempting to recover one or more of the three components of this process is known as the set of *inverse rendering* problems. The difference of the work proposed in this thesis from previous work is that we try to examine such problems through statistical frameworks that could potentially bridge the gap between this and higher-level computer vision tasks. This could ultimately allow the incorporation of such models of the components of image formation in larger frameworks towards the goal of scene understanding. To this end, we also attempt to relax the assumptions that similar approaches traditionally rely on. Relaxing these assumptions means that we are able to solve such inverse rendering problems for broader classes of images, including

complex natural images where knowledge of other components of the problem may be limited and unreliable.

The three components of the image formation process (geometry, reflectance, illumination) give rise to three different inverse rendering problems: shape reconstruction, reflectance estimation and illumination estimation. In this thesis we examine two of those problems. First we discuss illumination estimation from a single image, and present three approaches to model this problem in the case of approximate geometry knowledge, using the information contained in cast shadows. The third and more complete of these approaches combines ideas from the other two in a Markov Random Field formulation. This formulation is not only able to jointly estimate shadows and illumination parameters, but also able to incorporate information about geometry and estimate it jointly with the other two components. We demonstrate that the result is a framework that can be applied in complex natural images, using initial information that could be obtained automatically utilizing, for example, object detection.

Although through this approach we are able to infer geometry parameters that define the rough 3D geometry of occluders in the scene, the information contained in cast shadows is not adequate to estimate detailed 3D shape. We examine the problem of shape reconstruction in more detail, using shading variations as input. The problem of shape-from-shading has been a long-standing and challenging research area in computer vision. It is a generally ill-posed problem, with ambiguities that make a solution difficult. We propose a data-driven approach in an attempt to constrain these ambiguities, capturing priors on the local geometry directly through a dictionary of geometric primitives. The hypotheses produced by the dictionary to explain a test image are then combined through a graphical model.

The main contributions of this thesis are:

- An EM-based algorithm for illumination estimation, based on modeling illumination as a mixture of probability distributions, leading to a method that is robust to coarse geometry and that can model soft shadows (Chapter 4).

- The association of illumination directly with shadow edges instead of shadow intensities, which leads to a simple approach that is robust to geometry knowledge and shadow estimation (Chapter 5).

- The formulation of scene photometry, and specifically the cast shadow creation process, as a Markov Random Field model (Chapter 6). Such

a model offers robustness to inaccurate or incomplete information about geometry and shadows, as well as flexibility with regard to the cues used from illumination estimation. An efficient optimization scheme for this MRF is also proposed.

- The introduction of geometry parameters in the same MRF framework. This enables the joint, concurrent estimation of three major components of the problem of illumination estimation from shadows: cast shadows, illumination and geometry. In the same way higher-level information about objects in the scene could be introduced in our model, given the flexibility of the way geometry is parameterized in this model.

- A way to associate local shading patterns with the underlying local geometry, in a data-driven dictionary approach to the problem of shape reconstruction from shading (Chapter 7). An important contribution of our model is that it removes the Lambertian assumption by modeling the distribution of local shading patterns produced by a geometry patch over various reflectance parameter sets. This fact makes this approach able to get convincing shape reconstructions from real-world photographs.

The rest of this thesis is organized as follows:

Chapter 2 surveys the prior art in the four main areas of interest for this thesis: we present prior work in illumination estimation, with a closer look to illumination estimation from cast shadows; in cast shadow detection; in shape reconstruction from shading; and finally, in graphical models, with an emphasis on Markov Random Fields, since we will formulate our algorithms in the two main problems of interest through the use of such models.

Chapter 3 describes our approach to detecting cast shadow in images. The results from this approach are used as input to our method for illumination estimation through an MRF model in Chapter 6. The proposed shadow detection approach is based on simple observations about the nature of shadows in an image, and produces competitive results with low computational complexity.

Chapters 4, 5 and 6 present three approaches to illumination estimation from shadows. We choose cast shadows, instead of shading or specular highlights, as a cue for two reasons: on one hand, the cast shadows constitute the most stable among these three cues when knowledge of geometry is inaccurate or incomplete; specular highlights are the most heavily dependant on

accurate knowledge of the underlying geometry. On the other hand, cast shadows allow for estimation of the higher-frequency components of illumination (as discussed in [122]). This characteristic compares favorably to shading as an illumination cue, because the latter, in the case of lambertian reflectance, only allows for the estimation of low-frequency illumination components (since lambertian reflectance acts as a low-pass filter [8].

The first illumination estimation approach we examine, in Chapter 4, is based on modeling illumination as a mixture of distributions. This modeling provides certain advantages when applied to real world images; we propose an Expectation-Maximization (EM) algorithm to estimate illumination through this model [128]. We show how such an approach can allow the estimation of illumination in natural images, using 3D bounding boxes to model the geometry - a big step from the accurate 3D modeling assumed in past work. This approach has the extra advantage that it can model the perceived size of light sources, and therefore deal with soft shadows.

In Chapter 5 we propose a different approach to illumination estimation. Our approach is based on associating the light source parameters not with the pixel intensities in the image but with the observed image edges. In approaches that rely on the pixel intensity values, two significant types of errors can be introduced: errors in the initial shadow estimate propagate throughout the illumination estimation process, altering the final results; on the other hand, the knowledge of scene structure may not be adequate to explain a lot of correctly detected shadows in complex scenes, leading to erroneous illumination solutions that try to explain every observed shadow with inadequate geometry data. In this chapter we propose a way to couple shadow and illumination estimation, trying to detect only the shadow edges that are relevant to the provided geometry, as part of the illumination estimation process. This leads to an illumination estimation algorithm that can reliably estimate illumination, even when scene geometry knowledge is limited, while being less dependent on obtaining an initial shadow estimate (we can even avoid obtaining an estimate of shadow edges altogether, as we show in our results), and having lower computational complexity than state-of-the-art methods. In this approach, illumination estimation is posed as the minimization of an energy function, and coupled with the detection of salient shadow edges.

In chapter 6 we combine ideas from both previous approaches in a much more powerful framework. The proposed approach is based on formulating the creation of shadows in the image as a Markov Random Field. This statistical model provides not only robustness to rough geometry and initial shadow

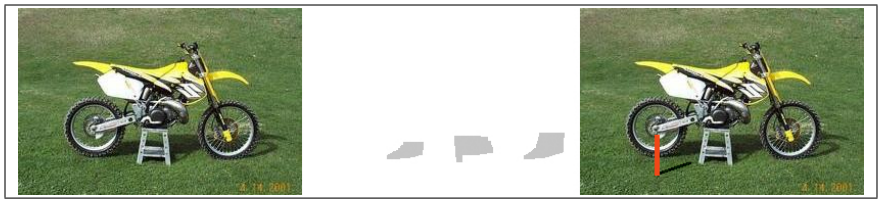

Figure 1.1: An example of our illumination estimation approach (Chapter 6): Left, the input image; center, the estimated shadow with our approach; right, the original image with a synthetic sundial (orange) rendered with the illumination estimate obtained with our approach. This is one example of the results obtained from estimating illumination in a large number of images from the "Motorcycles" class of Caltech101 [110] using the same coarse, average geometry and camera parameters for every instance.

information, but also allows us to directly incorporate more information about the scene, such as geometry parameters. The result is a model that enables the estimation of illumination in complex natural images when occluders are modeled by simple bounding boxes, or when a single approximate geometric model for a whole class of objects is used to approximate every object of that class. We further demonstrate how this model can enable inference of 3D geometry as a natural part of the illumination modeling. Through our framework, we are able to jointly estimate all three major components of our problem: the cast shadows, the illumination and the geometry parameters.

Chapter 7 shifts our focus to geometry, which could be dealt only in a coarse manner in the work of Chapter 6. In this chapter, we examine an approach to the problem of shape from shading based on the idea of learning shadow primitives. The goal of the work in this chapter is to infer the 3D scene structure, in the form of a normal map, from a single grayscale image using the information contained in shading. We capture the relationship between the appearance and geometry of image patches in a straight-forward way, by learning a dictionary that associates local image appearance with the underlying local geometry. The appearance is represented as a distribution of local appearances over different reflectances, to allow shape estimation even when surfaces deviate from the Lambertian assumption. When reconstructing the 3D shape of an image, we produce a set of local hypotheses about the geometry using the learned dictionary. These hypotheses are then combined in a Markov Random Field model in order to produce the final shape estimate.

Figure 1.2: An example of our 3D shape reconstruction approach (Chapter 7): Left, the original image [146]; center, the estimated normal map with our approach; right, a rendering of the estimated normal map under different illumination.

The primitives captured in the dictionary are effectively priors that constrain the ambiguities inherent in the shape-from-shading problem. As a result, we are able to demonstrate reliable shape reconstructions from both synthetic and real images, which can significantly outperform the state of the art, especially in the case of real photographs (see Fig.1.2 in this section for an example).

Finally, Chapter 8 concludes this thesis, summarizing the approaches proposed in previous chapters and discussing some directions for future research.

Chapter 2

Background Review

In this thesis we will examine two of the three instances of Inverse Rendering problems: estimating illumination from a single input image, and estimating 3D shape either jointly with illumination estimation, or directly from the observed shading patterns in the image. In this chapter we give an introduction to the fundamentals of these problems. We first examine the fundamentals of the problem of estimating illumination from one or more images, and discuss methods that have been proposed for this problem. We examine in more detail the case of estimation illumination from cast shadows, which will be the focus of subsequent chapters, and introduce some notation. We also discuss the literature on detecting shadows in an image, which is a problem interconnected with the estimation of illumination from shadows.

We then examine the literature in the second category of inverse rendering problems that is of interest: the problem of Shape-from-Shading (SfS). We give an overview of the large amount of literature that exists in the field, and a simple categorization of the proposed methods. We also examine the relatively sparse prior art that involves graphical models and data-driven approaches for this problem.

The approaches presented in this thesis differ with most of the approaches discussed in this chapter in that they try to incorporate knowledge about the inverse rendering problems in statistical frameworks, such as Markov Random Field models, that can both model the uncertainty in the input data and offer the flexibility to incorporate different cues and problem parameters. Therefore, after reviewing the methods proposed to solve this problem in the literature, we also give a brief introduction to graphical models, with an emphasis on Markov Random Fields. This introduction presents some fundamental concepts for the

discussion in Chapters 6 and 7.

2.1 Illumination Estimation

The problem of estimating illumination from one or more images is called *inverse lighting* [114]. One broad categorization of inverse lighting methods can be made based on the source of information utilized to estimate illumination. The three most common sources of information are specular reflections, shading and cast shadows.

Many techniques have been developed to estimate light source properties from shading variations in a single image, starting with the work of Pentland [137]. One directional light source can be estimated with the assumption that the viewed scene represents a convex object with sharp contours (Vega and Yang [183]; Yang and Yuille [193]). Yang and Yuille [193] analyze the intensities and surface normals along the occluding boundaries to estimate the directions of multiple light sources. Hougen and Ahuja [63] determine the light source directions and intensities from a single image of a Lambertian object of known geometry, solving a set of linear equations for image irradiance. Zheng and Chellappa [198] reconstruct the shape, illuminant direction, and texture from a single image of a Lambertian surface, using shading information along image contours. Marschner and Greenberg [114] propose a technique to reconstruct the directional distribution of light from an image, assuming Lambertian reflectance and accurate 3D geometry, by producing a set of basis images and finding a linear combination of those basis images that matches the input image. Kim et al. [72] estimate the illuminant direction from a single image of a Lambertian surface, while also recovering the shape of the surface, using image regions corresponding to bumps. Zhang and Yang [196, 197] detect critical points where the surface normal is perpendicular to some light source direction from a single image of a Lambertian sphere of known geometry and then determine the directions and intensities of multiple light sources. Wang and Samaras [187] extend that method by allowing Lambertian objects of arbitrary known shapes. This approach maps the surface normals onto a sphere and then segments the surface into regions, with each region illuminated by a different set of light sources. Finally, illuminant direction estimation is performed by a recursive least squares technique.

Specularities have also been used to estimate illumination parameters. Hara et al [55] propose a method to estimate surface reflectance and illu-

mination from a specular image, without the distant illumination assumption. Specularities are also utilized in [56] in order to estimate both illumination, in the form of multiple point light sources, and reflectance. Tominaga and Tanaka [176] utilize the dichromatic reflection model and the Phong model and successfully recovered the reflectance, light direction, and its color, texture, and shape under a single light source. Miyazaki et al. [117] present a simultaneous recovery of the shape, surface reflectance, texture, and the directions of multiple sources with polarization analysis of multiple images taken from a single view.

A directional light source can also be estimated with the additional help of shadows (Nillius and Eklundh [120], Sato et al. [166]). Sato et al. [165], [166] propose a method to simultaneously recover the illumination distribution (directions and intensities of light sources) and the surface reflectance by analyzing intensity information inside shadows cast on the scene by the object. The work on illumination based on shadows will be presented in more detail in the next section.

In other work, multiple cues are combined to reliably estimate the illumination distribution. Wang and Samaras [187] develop a method based on shadow and a method based on shading independently and integrate the two methods to estimate multiple directional illuminants. Li et al. [112] integrates cues from shading, shadow and specular reflections for estimating directional illumination in a textured scene. In [199], Zhou et al propose a unified framework to estimate both distant and point light sources. Other sources of information can also be used; [181] studies the problem of estimating illumination from images of textured surfaces, extending work in [76].

Illumination can also captured in a more direct way; in [25], Debevec proposes a method for acquiring a radiance map with photographs of a spherical surface mirror, such as a polished steel ball. Powel et al. [142] have also used specular spheres to estimate the position of several point light sources from a set of images.

Important theoretic results have been reported about the possible appearances of a diffuse object and their relationship with illumination. In [8] it is shown that the set of all reflectance functions (the mapping from surface normals to intensities) produced by Lambertian objects under distant, isotropic lighting lies close to a 9D linear subspace. [9] proves that the set of n-pixel images of a convex object with a Lambertian reflectance function, illuminated by an arbitrary number of point light sources at infinity, forms a convex polyhedral cone and that the dimension of this illumination cone equals the number

of distinct surface normals. [152], [150] develop a signal-processing framework which describes the reflected light field as a convolution of the lighting and BRDF, and expresses it mathematically as a product of spherical harmonic coefficients of the BRDF and the lighting. In [151] the subspace best approximating images of a convex Lambertian object under different illumination conditions is analyzed.

2.1.1 Illumination Estimation from Shadows

In this section, we present some background on the specific problem of illumination estimation from shadows in a single image, since this is a problem that will occupy a large part of this proposal.

In general, the outgoing radiance along direction ω at the 3D point $\mathbf{p}$ of the scene that projects to pixel i with coordinates (x, y) is

$$L_o(\mathbf{p}, \omega, \lambda) = \int_\Omega f_r(\mathbf{p}, \omega', \omega, \lambda) L_i(\mathbf{p}, \omega', \lambda)(-\omega' \cdot \mathbf{n_p}) d\omega', \tag{2.1}$$

where λ is the light wavelength, $f_r(\mathbf{p}, \omega', \omega, \lambda)$ is the BRDF, $L_i(\mathbf{p}, \omega', \lambda)$ is the incident radiance of wavelength λ at point $\mathbf{p}$ along direction ω', and Ω is the hemisphere of inward directions.

A commonly used set of assumptions is that the surfaces in the scene exhibit lambertian reflectance, and that the scene is illuminated by light sources at infinity, as well as some constant ambient illumination term. We discretize the integral in 2.1 using N sample directions on the illumination sphere. Under these assumptions, the outgoing radiance at a pixel i is given by:

$$L_o(\mathbf{p}) = \rho_\mathbf{p} \left(\alpha_0 + \sum_{i=1}^{N} V_\mathbf{p}(\mathbf{d}_i) \alpha_i \max\{\mathbf{d}_i \cdot \mathbf{n_p}, 0\} \right), \tag{2.2}$$

where $\rho_\mathbf{p}$ is the albedo at point $\mathbf{p}$, α_0 is the ambient intensity, $\alpha_i, i \in \{1, ..., N\}$ is the incoming radiance along the i-th sampling direction, $\mathbf{d}_i$ is the unit direction vector of the i-th sampling direction, and $V_\mathbf{p}(\mathbf{d}_i)$ is a visibility term for direction $\mathbf{d}_i$ at point $\mathbf{p}$, defined as:

$$V_\mathbf{p}(\mathbf{d}_j) = \begin{cases} 0, & \text{if ray to } \mathbf{p} \text{ along } \mathbf{d}_j \text{ intersects } \mathcal{G} \\ 1, & \text{otherwise} \end{cases} \tag{2.3}$$

Assuming a simplified linear model for the camera sensors, we model the

observed value at pixel (x, y) as:

$$I(x, y) = \kappa L_o(\mathbf{p}) + \epsilon, \tag{2.4}$$

where κ is an exposure parameter and ϵ is noise. Since we can only estimate light source intensities up to scale, we assume $\kappa = 1$.

Eq.2.2 taken for each image pixel (or a subset of image pixels) forms a linear system, with the illumination parameters as the unknowns. An important constraint in this system is the non-negativity of the light intensities:

$$\alpha_i \geq 0. \tag{2.5}$$

In [166] the set of illumination directions $\mathbf{d}_i$ is set to be an even sampling of all possible directions, corresponding to the nodes of a geodesic sphere. The albedo $\rho_{\mathbf{p}}$ is assumed uniform. In this case, the unknowns that need to be estimated for the inverse lighting problem are the intensity values $\{\alpha_i\}$ for each direction. This system in [166] is solved by non-negative least squares optimization.

This approach is however sensitive to inaccuracies in the 3D model that represents the scene, requiring 3D modeling that can be labor-intensive. Furthermore, it does not directly address the case of textured surfaces. An extra image containing the scene albedo is required in that case, in order to extract the shading from the input image.

[115] shows that the set of images produced by a Lambertian scene with cast shadows can be efficiently represented by a sparse set of images generated by directional light sources. They utilize this observation enforcing sparsity constraints in a linear system to better estimate illumination. First the image is separated in low-frequency, diffuse component and a high-frequency residual component which captures cast shadows. The low-frequency component, which mainly captures diffuse shading, is estimated using spherical harmonics. Then the high-frequency components are estimated solving a ℓ_1-regularized least-squares problem corresponding to a linear system with non-negativity and sparsity constraints.

Prior art on illumination estimation using shadows cast on textured surfaces is limited. In [166], an extra image is necessary to deal with texture. Li et al [111] propose a method that integrates multiple cues from shading, shadow, and specular reflections. Kim et al [73] use regularization by correlation to estimate illumination from shadows when texture is present, but requires ex-

tra user-specified information and assumes lambertian surface reflectance and known geometry.

Spherical Harmonics

The problem of illumination estimation from shadows can be alternatively formulated based on spherical harmonics. Spherical harmonics form an orthonormal basis defined over a unit sphere, where every square-integrable function can be projected. Spherical harmonics $Y_{lm}(\theta, \phi)$ are defined as:

$$Y_{lm}(\theta, \phi) = N_{lm} P_l^m(\cos\theta) e^{im\phi}, \tag{2.6}$$

where $P_l^m(\cdot)$ are the associated Legendre functions and N_{lm} the normalization constants.

Since we assume distant illumination, the illumination function is defined on a sphere. We can expand the illumination L on spherical harmonics $Y_{lm}(\theta, \phi)$ using spherical coordinates $\omega = (\theta, \phi)$, obtaining

$$L(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} L_{lm} Y_{lm}(\theta, \phi). \tag{2.7}$$

Therefore, on the spherical harmonic basis, illumination is defined by coefficients L_{lm}. These coefficients are the variables to be estimated in order to estimate illumination. In order to render the value of an image pixel i (corresponding to 3D point p) using spherical harmonics, we need to integrate incoming illumination L and a transfer function t:

$$I(i) = \int_{S^2} L(\omega) t(\omega, \mathbf{p}) d\omega. \tag{2.8}$$

If we project both the illumination and the transfer function into spherical harmonic coefficients, the integral becomes a simple dot product:

$$I(i) = \sum_{k=0}^{n^2} L_k t_k(\mathbf{p}), \tag{2.9}$$

where L_k and $t_k(\mathbf{p})$ are the spherical harmonic coefficients of illumination L and the transfer function t at point $\mathbf{p}$ respectively. In the above equation,

we have kept only the first $n+1$ harmonic coefficients, making the equality relationship approximate.

If we assume lambertian reflectance, and taking into account shadows, the transfer function can be directly calculated by the scene geometry $\mathcal{G}$:

$$t(\omega, \mathbf{p}) = V_{\mathbf{p}}(\omega) \max(\mathbf{n}(\mathbf{p}) \cdot \omega, 0), \tag{2.10}$$

following Eq.2.2. $V_{\mathbf{p}}(\omega)$ is a binary visibility term, as in Eq.2.3.

The problem of illumination estimation in this setting corresponds to the estimation of the spherical harmonic coefficients of illumination L_k in Eq.2.9, where B is the observed image and coefficients t_k can be computed from the geometry. There is one equation for each image pixel, forming a simple linear system

$$\mathbf{I} = \mathbf{t}\mathbf{L}^T, \tag{2.11}$$

where $\mathbf{I} = [\{I(i)\}]$ is a known vector of size N, $\mathbf{t}$ is a known matrix of size $N \times (n+1)^2$, $\mathbf{L}$ is the unknown vector of size $(n+1)^2$ and N is the number of pixels in the image.

The relationship of illumination and shadows is examined more closely in [122], where it is shown that the transfer function t has non-zero high-frequency components when cast shadows are taken into account. One important result of this observation is that high-frequency components of illumination contribute to the brightness of the surface (under lambertian reflectance), making it possible to use cast shadows to estimate such high-frequency illumination components. Such high-frequency components cannot be estimated by the shading on lambertian surfaces [150], as mentioned earlier.

Estimating illumination from shadows using spherical harmonics has however certain limitations:

- First, a very large number of basis functions is required to estimate illumination that is well localized in the angular domain, such as a point light source
- Second, high-frequency components are harder to estimate as the number of shadow pixels observed is reduced, making the estimation of high-frequency components often difficult.

In [122] an improvement is proposed by using Haar wavelets as the basis functions. This offers the advantages of basis functions with compact supports

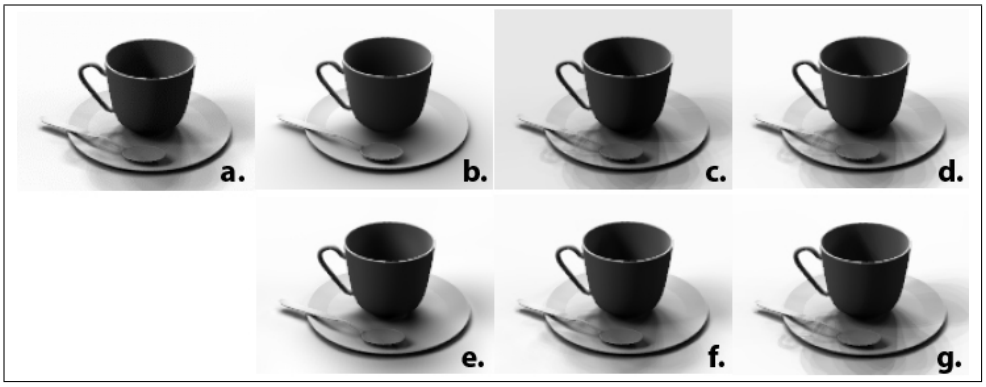

Figure 2.1: A comparison of different methods for illumination estimation from shadows. a) Original image; b) 3D model rendered with illumination estimated using spherical harmonics; c) result with [166] (100 directions); d) result with [166] (300 directions); e) result with [168]; f) result with [122]; g) result with [115]. Images from [115].

and sparsity. The latter means that many of the resulting illumination coefficients are near-zero, and fewer that 1% of the basis functions are sufficient to accurately represent natural illumination [119].

A Fourier analysis of cast shadows for the case of surfaces that exhibit 3D texture with canonical configurations, such as V-grooves, is examined in [153].

In [168] a spherical harmonic representation is used in combination with semidefinite programming to estimate illumination under non-negativity constraints. While [168] focuses on specular objects, the same technique applies to the case of cast shadows as well.

Illumination Estimation with Limited Geometry Knowledge

The work discussed so far has assumed that geometry is known accurately. Much fewer methods have tackled the problem of inverse lighting when the knowledge of 3D scene geometry is limited and inaccurate.

Recently Lalonde et al [97] proposed an approach that combines cues from the sky, cast shadows on the ground and surface brightness to estimate illumination of outdoor scenes with the sun as the single light source. Their method makes strong assumptions and is only applicable to daytime outdoor scenes. Karsch et al [71] utilize simplified geometry provided through user annotation to estimate various components of a scene, including illumination.

Their approach can convincingly insert synthetic objects in real scenes using the obtained estimates of the various scene components.

The ability to estimate illumination with limited knowledge of geometry significantly increases the practicality and usefulness of inverse lighting. We will extensively deal with this case in later chapters.

2.2 Shadow Detection

As mentioned in the previous sections, in this thesis we will examine more closely the estimation of illumination from the cast shadows in an image. This assumes that we are able to detect cast shadows in the first place. Cast shadow detection from a single image is however a difficult problem in the general case. When shadows are cast on textured surfaces and in general, complex scenes captured in low dynamic range images, their detection can be challenging. Hence in this section we review prior work in the problem of detecting cast shadows.

The detection of cast shadows in the general case is not straightforward. Shadow detection, in the absence of illumination estimation or knowledge of 3D geometry is a well studied problem. [161] uses invariant color features to segment cast shadows in still or moving images. [109] suggests a method to detect and remove shadows based on the properties of shadow boundaries in the image. In [32, 33], a set of illumination invariant features is proposed to detect and remove shadows from a single image. This method is suited to images with relatively sharp shadows and makes some assumptions about the lights and the camera. Camera calibration is necessary; if this is not possible, an entropy minimization method is proposed to recover the most probable illumination invariant image. In [169], a method for high-quality shadow detection and removal is discussed. The method, however, needs some very limited user input. Recently, [200] proposed a method to detect shadows in the case of monochromatic images, based on a number of features that capture statistical properties of the shadows. Lalonde et al [98] propose a learning approach to detect shadows in consumer-grade photographs, focusing on shadows on the ground. The above methods detect the majority of shadow pixels, but they are not always accurate since they are based only on image statistics.

A related body of work involves the extraction of *intrinsic images* from the input image [38, 172, 173, 46]. The intrinsic images can separate the

albedo and the shading from the input image, based on learned image statistics. However, they do not target the identification of the cast shadows specifically, and separately from shading.

In [130] a bayesian framework is proposed for shadow extraction from a single image, with no assumptions about the camera and lights, and assuming Lambertian reflectance. This method requires, however, information supplied by the user in the form of a rough *quadmap* which approximately identifies shadow and non-shadow regions.

2.2.1 Illumination Invariants

Photometric color invariants are functions which describe each image point, while disregarding shading and shadows. These functions are demonstrated to be invariant to a change in the imaging conditions, such as viewing direction, object's surface orientation and illumination conditions. Some examples of photometric invariant color features are normalized RGB, hue, saturation, $c_1c_2c_3$ and $l_1l_2l_3$ [41]. An examination of various photometric invariants is given in detail in [42]. Other interesting invariants that could be exploited are described in [40, 179, 27].

Hue and saturation are two simple invariants; both hue and saturation are shown to be invariant to surface orientation, illumination orientation and illumination intensity, while hue is also invariant to specular highlights [42].

The normalized RGB color invariant is defined as:

$$r = \frac{R}{R+G+B}, \tag{2.12}$$

$$g = \frac{G}{R+G+B}, \tag{2.13}$$

$$b = \frac{B}{R+G+B}, \tag{2.14}$$

which is shown [42] to be insensitive to surface orientation, illumination direction and illumination intensity; R, G and B are the three components of RGB color.

The $c_1c_2c_3$ invariant color feature is defined as:

$$c_1 = \arctan\left(\frac{R}{max\{G,B\}}\right), \tag{2.15}$$

$$c_2 = \arctan\left(\frac{G}{max\{R,B\}}\right), \tag{2.16}$$

$$c_3 = \arctan\left(\frac{B}{max\{R,G\}}\right), \tag{2.17}$$

where R, G and B are the three components of RGB color. It is shown [42] to be invariant to surface orientation, illumination direction and illumination intensity.

Multiple invariants can be combined in a tensor framework to offer more robust illumination-invariant edge detection [42].

Adding to the various illumination invariants, a set of photometric quasi-invariants has been proposed [180, 27]. The various quasi-invariants are also insensitive to certain photometric edges, such as shadows and shading. They are advantageous for the task of feature detection invariant to illumination effects, since they don't exhibit the non-linear nature of photometric invariants. The non-linear nature of the latter means that they lead to unstable features. Quasi-invariants can be chosen in order to isolate the illumination-invariant edges in an image, or, alternatively, the edges attributed to specific photometric features such as shadows and shading [42].

Another illumination invariant representation specifically targeted to shadows is described in [33]. For this representation, a vector of illuminant variation e is estimated. The illumination invariant features are defined as the projection of the log-chromaticity vector x' of the pixel color with respect to color channel p to a vector $e^{\perp}$ orthogonal to e:

$$I' = \mathbf{x}'^T e^{\perp} \tag{2.18}$$

$$x'_j = \frac{\rho_k}{\rho_p}, k \in 1,2,3, k \neq p, j = 1,2 \tag{2.19}$$

and ρ_k represents the k-th RGB component.

The illumination invariant features of [33] assume narrow-band camera sensors, Planckian illuminants and a known sensor response, which requires calibration. The known sensor response requirement can be circumvented by using the entropy-minimization procedure proposed in [32] to calculate

the illuminant variation direction e. Futhermore, it has been shown that the features extracted this way are sufficiently illumination-invariant, even if the other two assumptions above are not met ([33]). A weakness of this feature in practice is that its performance decreases in the case of images that have been degraded, for example by JPEG compression, a case often true in practice.

2.2.2 Combining multiple cues

More recently, and concurrently with work presented in this thesis, some new approaches have appeared that combine different cues in statistical frameworks in order to detect shadows.

Lalonde et al [98] propose an algorithm to automatically detect shadows cast by objects onto the ground, putting an emphasis on outdoors consumer-grade photographs where approaches such as [33] are not effective. They only examine shadows cast onto the ground. Their key idea is that the appearance of the ground in outdoor images falls into a small number of predefined classes, corresponding to common materials/textures such as stone, grass, asphalt etc. They aim to learn the appearance of this limited number of classes from a labeled training set in order to recognize shadows. Their approach has three components: a) using existing ground classifiers to recognize the portion of an image that corresponds to ground; b) training a decision tree classifier on shadow-specific features around image edges; c) and a Conditional Random Field model (CRF) that combines the shadow edge detection results to create coherent shadow contours. Their detection accuracy is around 85% but they are limited to shadows cast on the ground in daytime outdoor images.

Zhu and Tappen [200] focus on the more difficult problem of detecting shadows in grayscale images. They use both shadow-variant and shadow-invariant cues from illumination, textural and odd-order derivative characteristics. The features they employ, based on an oversegmentation of the image, include intensity difference, the local maximum intensity smoothness, the skewness of intensities distribution, gradient and texture similarity features, entropy and the sum of edge responses. A boosted decision tree classifier is trained on this large number of features and then integrated into a CRF model. The CRF enforces local consistency over the pixel shadow labels. Their approach achieves detection rates of almost 89% in a dataset of challenging grayscale natural images. It depends however on training and is computationally intensive.

Guo et al [47] propose a region based approach. In addition to considering individual regions separately, they train a pairwise classifier to predict

illumination conditions between segmented regions from their appearances. They combine the single-region and pairwise classifiers in a graphical model corresponding to the image segments, and employ graph-cuts to find a labeling of shadow and non-shadow regions. They finally refine their results using image matting, in order to also be able to remove shadows. Their approach achieves slightly higher classification results compared to the previous approaches, reaching up to 90% classification rates in the same dataset as [200].

2.3 Shape Recovery

In this section we review the prior art relevant to the second inverse rendering problem that we will examine in later chapters, that of shape recovery. Shape recovery is a classic problem in computer vision and a large body of prior work exists on the subject, including a variety of shape-from-X techniques. The goal of shape recovery methods it to infer the 3D geometry underlying a scene from one or more images of that scene.

The reconstructed 3D geometry can be expressed in different ways:

- Depth values $z(\mathbf{x})$ at point $\mathbf{x}$, which can be measured either as the distance of surface points from the camera, or as the height from the $x-y$ plane.
- Surface normals $\mathbf{n} = (n_x, n_y, n_z)$, which are vectors perpendicular to the tangent plane on the object surface.
- Surface gradients $(p, q) = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right)$, which are the rate of change of the surface depth along the x and y directions. The surface normal can be related to the surface gradients by $\mathbf{n} = \frac{1}{\sqrt{1+p^2+q^2}}(-p, -q, 1)$.
- Slant, ϕ, and tilt, θ, angles.These angles define the surface normal in a spherical coordinate system. The surface normal is expressed in terms of ϕ and θ as $\mathbf{n} = (\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi)$.

Shape-from-shading is the instance of the shape recovery problem where shape is inferred by the variations of the shading variations, observed as the image *brightness*, in *a single* image. Although shading is a very important cue for human perception of shape and depth, shape-from-shading is a challenging and generally ill-posed problem in computer vision.

2.3.1 Shape from Shading

Shading is an important cue for the perception of shape by the human visual system. In humans, information from shading is combined with stereoscopic processing, the information from outlines, elementary features and the visual system's knowledge of objects [148, 7], in order to rapidly and accurately perceive the 3D shape of surfaces. In computer vision, the shape-from-shading problem (SfS) has mostly focused on a set of simplifying assumptions. These assumptions, including Lambertian reflectance and known light source direction, are not necessarily valid for the human visual system [116], but have been widely used in order to find tractable solutions to the problem. Despite the large amount of prior art, the applicability of shape-from-shading methods in practice has been limited. The best results are obtained when combined with some other strong prior, such as stereo reconstruction (e.g. [163]).

Horn [60, 61] in the 70's was the first to formulate the Shape From Shading problem rigorously as that of finding the solution of a nonlinear first-order Partial Differential Equation (PDE), the "image irradiance equation":

$$R\left(\mathbf{n}\left(\mathbf{x}\right)\right) = I(\mathbf{x}), \tag{2.20}$$

where $I(\mathbf{x})$ is the image brightness at point $\mathbf{x}$ (proportional to the image irradiance), $\mathbf{n}(\mathbf{x})$ is the normal vector at $\mathbf{x}$ and $R\left(\mathbf{n}\left(\mathbf{x}\right)\right)$ is the reflectance function which gives us the radiance at point $\mathbf{x}$ as a function of the normal $\mathbf{n}(\mathbf{x})$.

Let $z(\mathbf{x})$ be the height of the surface. Here we assume that the scene is illuminated by a single known point light source at infinity, whose direction is $\mathbf{d} = (d_x, d_y, d_z)$ and that the surface has uniform albedo equal to 1 and exhibits Lambertian reflectance. Then we can re-write Eq.2.20 as:

$$I(\mathbf{x}) = \sqrt{1 + \left|\nabla z(\mathbf{x})\right|^2} + (d_x, d_y)\nabla z(\mathbf{x}) - d_z = 0, \tag{2.21}$$

which is a first-order non-linear Hamilton-Jacobi PDE. If we assume that the light source is a frontal light source at infinity, so that $\mathbf{d} = (0, 0, 1)$, then Eq.2.21 becomes the *eikonal equation*:

$$\left|\nabla z(\mathbf{x})\right|^2 = \sqrt{\frac{1}{I(\mathbf{x})^2} - 1}. \tag{2.22}$$

Figure 2.2: Examples of ambiguities in shape-from-shading. Left: the crater illusion (from [135]). If one imagines that the light source is at the bottom of the image, the two craters can be thought of as two upside down volcanoes (the image actually depicts two ash cones). Right: the bas-relief ambiguity in a marble bas-relief sculpture (from [10]). While from the frontal view we expect the sculpture to have full 3D depth, the actual surface is the "flattened" surface shown in the side view.

For a long time, research in shape-from-shading focused on the computational part of the problem, trying to directly compute numerical solutions. Soon, however, and due to the poor quality of the results, questions about the existence and uniqueness of solutions became central. The shape-from-shading is now known to be an ill-posed problem [20, 124, 126, 10]. One common type of ambiguities is that between convex and concave surfaces that can produce the same image brightness (see Fig.2.2 for an example). When the light source and surface albedo are unknown, Belheumer et al [10] proved that the same image can be obtained by a continuous family of surfaces. This is known as the "Bas-relief ambiuity [10]. Therefore, in general cases, we cannot unambiguously infer the 3D structure of an object, as seen for a single viewpoint, using the shading and shadowing.

A lot of research has focused on the above set of assumptions, with results of limited applicability. An interesting modification to the above assumptions is the replacement of the orthographic projection with the more realistic perspective projection [133, 162]. The equations for perspective shape-from-shading are established by [143, 145, 22], and it is also a nonlinear PDE. Another modification to the "classical" assumptions, that leads to interesting changes to the problem formulation, is the replacement of the light source at infinity with a light source at the center of projection [123, 146]. These two modifications to the problem assumptions can lead to a formulation that is not ill-posed [146].

A third important assumption above has been that of Lambertian reflectance. Very little work has dealt with surfaces of non-Lambertian re-

flectance [3, 104, 147].

The work in shape-from-shading can be divided in the following broad categories:

- Local approaches: Local approaches for shape-from-shading involve examining small image regions and reconstructing the local shape. These local results are then quilted together. Local approaches tend to be fast, but often require some initial information about the surface, such as the linearization of the reflectance map [135, 136] or the depth at singular points [124]. In practice, when applied to real-world images such techniques have not been effective.

- Global approaches: global approaches recover the entire 3D surface, either through propagation of height values from singular points of the surface, or by minimizing some energy functional defined on a parametrization of the reconstructed 3D surface.

  - Global propagation techniques propagate information from specific points of the 3D surface. Examples of such solutions are, for example, various approaches based on viscosity solutions [159].
  - Global minimization approaches express constraints on the solution in the form of an energy function. Minimizing this energy recovers a surface that attempts to satisfy such constraints. Minimization approaches have been shown in the past to be more generally applicable to different types of input images, and more robust to noise than either local techniques or global propagation approaches [195].

Local Approaches

Pentland [135] assumed that surfaces are locally spherical at each point. He reconstructed 3D shape from the image intensity and its first and second derivatives. Lee and Rosenfeld [102] used the same assumption of locally spherical surfaces to compute the slant and tilt of the surface in the light source coordinate system based on the first derivative of image intensity. Pentland [134] used the linear approximation of the reflectance function in terms of the surface gradient, and applied a Fourier transform to the linear function to get a closed form solution for the depth at each point. Tsai and Shah [177] applied the discrete approximation of the gradient first, then employed the linear ap-

proximation of the reflectance function in terms of the depth directly. Their algorithm recovered the depth at each point using a Jacobi iterative scheme.

Global Propagation Approaches

Horn's original method [59] based on characteristic strips can be categorized as a propagation method. A characteristic strip is a line in the image, along which the surface depth and orientation can be computed if these quantities are known at the starting point of the line. This method constructs initial estimates around the singular points in the shading image using a spherical approximation. Characteristic strips are assumed to have the same direction as intensity gradients. The initial shape information is propagated along the characteristic strips.

A category of methods that can be categorized as global propagation approaches are those based on viscosity solutions. The notion of viscosity solutions was first used to solve SFS problems by Rouy and Tourin [159]. Their work is based on the notion of continuous viscosity solution, which are PDE solutions that may not be differentiable and may contain edges. In their work, they provide conditions for the existence of both continuous and smooth solutions and a numerical scheme to obtain a solution based on dynamic programming. Oliensis [124] described how the surface shape can be reconstructed from singular points instead of the occluding boundary. Based on this idea, Dupuis and Oliensis [127, 125] formulated SfS as an optimal control problem, and solved it using numerical methods. Bichsel and Pentland [13] simplified Dupuis and Oliensis's approach and proposed a minimum downhill approach for SFS which converged in less than ten iterations. Similar to Horn's, and Dupuis and Oliensis's approaches, Kimmel and Bruckstein [75, 74] reconstructed the surface through layers of equal height contours from an initial closed curve. Their method applied techniques in differential geometry, fluid dynamics, and numerical analysis, which enabled the recovery of non-smooth surfaces. The algorithm used a closed curve in the areas of singular points for initialization.

The perspective projection, in combination with the assumption of a light source at the center of projection, is examined by Prados and Faugeras [146]. They are able to show that under such assumptions, the problem of shape from shading can be well-posed. Based on the notion of viscosity solutions, they provide [145, 144] a generic, provably convergent shape-from-shading method applicable to both orthographic and perspective projection.

Global Minimization Approaches

Ikeuchi and Horn [64] propose one of the earliest minimization approaches to recover the surface gradients. The surface gradients are defined by two values for each pixel, constrained only by a single intensity value. To solve the resulting underdetermined system, the introduce a smoothness constraint that requires the result of the reconstruction to be a smooth surface. Surface gradients are also constrained by the brightness constraint, requiring the produced brightness to be the same as the observed. An energy function that expresses the above two constraints is minimized to reach a solution. To ensure convergence, initial knowledge of the shape at the occluding boundaries is required. Brooks and Horn [21] minimize an energy function expressing these two constraints in terms of the surface normals.

Frankot and Chellappa [35] enforce integrability in the surface slopes reconstructed with algorithms such as [21]. The *integrability constaint* is an important constraint for SfS problems, requiring the reconstructed normal maps or gradient fields to correspond to plausible 3D surfaces. In Frankot and Chellapa's approach, a possibly nonintegrable estimate of surface slopes is represented by a finite set of basis functions, and integrability is enforced by calculating the orthogonal projection onto a vector subspace spanning the set of integrable slopes. They also examine the special case of Fourier basis function, leading to a frequency domain interpretation of shape-from-shading. With this approach they are able to improve both accuracy and efficiency over Brooks and Horn's algorithm [21]. Subsequently, Horn also replaces the smoothness constraint in his approach with an integrability constraint [62]. One issue with Horn's method is slow convergence. Szeliski [170] proposes a hierarchical basis pre-conditioned conjugate gradient descent algorithm to improve computational efficiency. Vega and Yang [184] propose a heuristics-based SfS approach (the shading logic algorithm). They derive the heuristics from the geometric interpretation of Brooks and Horn's algorithm [21] in order to improve the performance and stability of Brooks and Horn's algorithm. Zheng and Chellappa [198] introduce an intensity gradient constraint, instead of the more common smoothness constraint. This constraint requires the intensity gradients of the reconstructed image and the input image to be close. An overview of various constraints proposed for the consistency of reconstructed normal maps is given in [192].

The above techniques are based on variational calculus. A discrete formualtion was used by Leclerc and Bobick [101], who solved for depth values

using a conjugate gradient technique. Their formulation is also constrained by the brightness and smoothness constraints. Initialization is required through the use of stereo reconstruction. Lee and Kuo [103] do not require an initial depth estimate, and model the surface using triangular patches. The solution also attempts to satisfy the brightness and smoothness constraints. Relaxing the assumption of a single smooth surface, Malik and Maydan [113] assumed piecewise smooth surfaces. They reconstruct shape by minimizing an energy function that combines constraints both from shading and line drawing. Energy minimization then recovers both a normal map and a line labeling.

The introduction of shading constraints in a physics-based deformable model framework is first examined by Samaras et al. [162, 164]. In this work, they provide a general methodology for the incorporation of illumination constraints within a deformable model framework and apply it to the coupled problems of shape from shading and light source estimation from images. Their method can incorporate any type of shading constraint, from Lambertian to highly non-linear ones, and can be applied to both orthographic or perspective projection. Potetz [138] formulates the shape-from-shading problem as a Markov Random Field. This formulation results in a higher-order MRF model, that incorporates integrability constraints and simple priors about the 3D shape. Energy minimization on this model correspond to a MAP estimate of the surface gradients. Although they focus on the efficient optimization of such higher-order MRF models, the approach is still very computationally expensive.Recently, Barron and Malik [6] proposed a method to solve simultaneously the problems of shape-from-shading and intrinsic images, by imposing "naturalness" priors over albedo and shape. Their approach is aided by an initial low-frequency estimate of the 3D shape.

2.3.2 Local Shading Patterns

Later in this thesis, in Chapter 7, we are going to describe a data-driven approach to shape-from-shading that utilizes information in larger image regions (*image patches*) consisting of many pixels. The work we examined so far generally aims to constrain the 3D surface at a single image pixel based on the observed image intensity and its immediate neighbors, using simple constraints such as smoothness and assumptions about the reflectance model.

Some prior work has dealt with the relationships between shading and geometry in small image regions. Haddon and Forsyth [49, 50] notice that the effect of interreflections due to distant surfaces is confined to low spatial fre-

quencies in the shading field. They look for "stereotyped" appearance patterns in the shading field that are linked to specific geometric primitives, such as folds and grooves. In such patterns, the effect of unseen surfaces in the environment has effects that vary slowly over the region of support. They examine approaches for testing hypotheses of geometric primitives for consistency with the shading field, and looking for shading patterns that are distinctive of some shape pattern. These approaches can be composed into a bottom-up process of representation. Han et al [54] target the specific case of folds in cloth, and learn a set of shading primitives to represent them. Through these primitives they are able to reconstruct the geometry of folds, and the surface in between folds is interpolated through a two-level MRF model to get the complete 3D shape. Recently, Varol et al [182] proposed learned shading primitives to deform the initially known 3D surface of a locally textured object. They learn a mapping between local shading patterns and the deformations of the underlying shape. They use this mapping to reconstruct parts of deformable surfaces that are not well-textured, as part of a nonrigid shape recovery approach driven by point correspondences in video sequences.

The relationship between shading and geometry patterns has also been examined through the use of neural networks. Lehky and Sejnowski [106] show that it is possible to reconstruct the approximate surface normals in the case of simple ellipsoidal shapes by training a neural network on intensity patterns. The intensity variations in image patches are also examined in [189], where a multilayer feedforward network is applied to the SfS problem. Similarly, [11] proposes a backpropagation-based neural network for learning brightness patterns and associating them with range data. In [140], [139] the statistical relationship between 2D appearance and the underlying 3D geometry is examined. Lee et al [105] suggest neural network configurations that could simulate how the human brain takes advantage of such statistical properties to infer information about surfaces, which could lead to useful priors for statistical 3D surface inference in computer vision.

2.4 Graphical Models

Most of the approaches in the literature we surveyed in the previous sections are based on directly modeling the physical properties of the problem and attempting to find numerical solutions to the resulting systems of equations. In order to describe the physical laws governing the problems of interest, usually

these methods have to rely on several strong assumptions. When such assumptions are violated, the quality of the results obtained will suffer. It is, however, to be expected that assumptions such as Lambertian reflectance or knowledge of geometry will be violated when examining a real, complex image. In such cases, the assumptions we make can be at best only satisfied approximately. It is the goal of the work in this thesis to extend the applicability of solutions to inverse rendering problems to such real, complex images. To this goal, we are interested in modeling the problems in ways that, on one hand, can model the uncertainty in our data, and on the other, provide flexible frameworks that are able to incorporate a variety of features and parameters that are relevant to each application. An important tool we will use is *graphical models*, hence this section gives an introduction to the fundamentals behind them.

Graphical models offer a powerful framework to express the statistical dependencies in a large variety of problems across different scientific or engineering fields. They have become popular in computer vision problems due to the following properties, which motivated us to also examine their use in the aforementioned inverse rendering problems:

1. Graphical models lead to flexible, modular frameworks. The graph structure, and the corresponding natural factorization of the probability distribution they model, makes it natural to add and remove "components" that model subsets of the problem. Therefore, a graphical model framework can often be easily extended to incorporate different sources of information in the form of input data or priors. We take advantage of this property to allow for different parametrizations of geometry in Chapter 6.

2. The existence of inference methods applicable to large classes of MRF models allows the decoupling of the modeling and inference. This allows easier modeling of the problem, and facilitates the combination of components in larger frameworks. Furthermore, discrete optimization in graphical models enables tractable inference in a variety of difficult problems, relaxing many of the constraints posed on the energy function by continuous methods. This has lead to very large range of applications of graphical models in computer vision problems.

3. The probabilistic nature of graphical models has potential advantages, compared to classic variational methods, in terms of parameter learning [155, 160] and uncertainty analysis [83, 44].

Graphical models have been widely used in Computer Vision problems (image restoration, image segmentation, stereo reconstruction etc.), where they have offered a more natural way to formalize prior knowledge about the structure of each problem, and allowed inference in the complex models that arise as a result. Markov Random Fields (MRFs), in particular, have become a ubiquitous tool in computer vision problems.

A graphical model is a probabilistic model which corresponds to a graph. That graph denotes the conditional independence between the random variables of the model. More specifically, each graphical model can be represented by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, consisting of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$. By $\mathbf{X} = \{X_i\}_{i \in \mathcal{V}}$ we denote a set of random variables indexed by the nodes $\mathcal{V}$, so that for each node $i \in \mathcal{V}$, there is an associated random variable X_i. By x_i we denote a realization of the random variable X_i, taking values from a state space $\mathcal{X}_i$. We will often refer to such a realization of random variable X_i as a *label* of node i.

The graphical model provides a compact representation of a family of joint distributions over the multidimensional space formed by the Cartesian product of the state spaces $\mathcal{X}_i$ of each random variable X_i. The lack of an edge between two nodes i and j denotes conditional independence between the corresponding random variables X_i and X_j in this family of joint distributions. Most joint distributions of interest represented by a graphical model can be factorized into a product of local functions, each of which involves a (usually small) subset of random variables. This factorization is a key concept behind graphical models.

Two types of graphical models are commonly used:

- *Bayesian networks*, which are represented by directed acyclic graphs (also known as *Belief Networks*)
- *Markov networks*, which are represented by undirected graphs (also known as *Markov Random Fields*)

Each of these model types can represent certain dependencies that the other one cannot: A Markov network can represent cyclic dependencies which a Bayesian network can't; the latter can represent induced dependencies which the former cannot. Both types of models can be represented by a unified representation called a *factor graph* (if the corresponding joint distribution can be factorized).

The interested reader can refer to [100, 14, 69, 84] for a more in-depth analysis. In the next sections we describe in more detail one graphical model

type, the Markov Random Field, and give a brief overview of methods to perform inference on such models.

2.4.1 Markov Random Fields

Markov Random Field (MRF) models have been widely used in Computer Vision to model various low- or mid-level tasks. Examples of applications of MRFs in Computer Vision are image denoising [12, 155], image super-resolution [37], segmentation [167], stereo-matching [30] etc.

A Markov Random Field is a set of variables having a Markov property described by an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. Formally, to form a Markov random field, the set of random variables $\mathbf{X}$ indexed by $\mathcal{V}$ must satisfy the following local independence assumption, called *local Markov property*:

$$\forall i \in \mathcal{V}, X_i \perp X_{\mathcal{V}-\{i\}} | X_{\mathcal{N}_i}, \tag{2.23}$$

which means that each node i is independent of all other nodes given all of its neighbors $\mathcal{N}_i$.

The associated family of joint distributions $p(\mathbf{x})$ (satisfying the local Markov property) are *Gibbs distributions*. Given the graph $\mathcal{G}$ with cliques $\mathcal{C}$, the joint distribution can be factorized in the form:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c), \tag{2.24}$$

where $\phi_c(x_c)$ is a *potential function* defined over clique c. Potential function $\phi_c(x_c)$ takes positive real values over the possible configurations x_c of clique c. Z is a normalizing factor such that $p(x)$ is a probability distribution.

We can define the *MRF energy* $E(\mathbf{x})$ as the sum of all potential functions:

$$E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \psi_c(x_c), \tag{2.25}$$

where

$$\psi_c(x_c) = -\log \psi_c(x_c) \tag{2.26}$$

is the clique energy, which is also referred to as a *clique potential* or potential

function. The joint distribution $p(\mathbf{x})$ can then be written as:

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{-E(\mathbf{x})\right\}. \tag{2.27}$$

Inference of the most probable configuration $\mathbf{x}^{opt}$ of variables $\mathbf{X}$ be performed by *maximum a posteriori (MAP)* estimation:

$$\mathbf{x}^{opt} = \arg\max_{\mathbf{X}} p(\mathbf{x}). \tag{2.28}$$

From the above, it can be seen that MAP inference corresponds to minimizing the MRF energy $E(\mathbf{x})$:

$$\mathbf{x}^{opt} = \arg\min_{\mathbf{X}} E(\mathbf{x}), \tag{2.29}$$

2.4.2 Inference on MRF Models

Many important computer vision problems can be elegantly expressed in terms of MAP estimation of a Markov Random Field. However, despite the power of the energy minimization approach to such problems, early attempts were limited by computational considerations. In particular, the algorithms originally used for energy minimization in MRFs, such as *iterated conditional modes* (ICM) [12] or *simulated annealing* [5] appeared to be very inefficient. Several related classes of energy minimization problems were viewed as intractable. However, in more recent years, powerful methods such as *graph cuts* [24, 18, 16] and *loopy belief propagation* [194] were popularized and were proven to be very powerful [171]. Furthermore, significant effort has been dedicated recently to developing efficient energy minimization approaches for even more challenging MRF classes, such as models that contain *higher-order cliques* (cliques of order higher than 2). Such developments are significant, because they greatly extend the range of MRF models that can be used in practice in order to include many useful cases. In this section, we gave a brief overview of the developments in MRF inference, both for the more standard MRF topology that only includes pairwise interactions, as well as for the more challenging case of higher-order graphs.

Graph Cut Methods

Graph cuts was originally developed [24] for MRFs where labels x_i take binary values. The main idea behind graph cuts is to introduce two special nodes s

and t called the *source* and *sink*. A directed graph $\mathcal{G}^{st}$ is constructed to include the source and sink. An s-t cut partitions the nodes of this graph into to disjoint sets S and T, so that $s \in S$ and $t \in T$. For each directed edge i, j, a non-negative capacity setting $c(i, j)$ is assigned, so that the cost $C(S, T)$ of the cut is equal to the MRF energy of the corresponding (binary) configuration $\mathbf{x}$. If an MRF has such a graph representation, it is called *graph-representable*. The minimization of the energy of an MRF that can be represented in this form corresponds to minimizing the cost of the s-t cut (min-cut problem):

$$C(S, T) = \sum_{i \in S, j \in T} c(i, j), \tag{2.30}$$

which is equivalent to a max-flow problem and can be computed in polynomial time. However, not all MRFs are graph-representable. It has been shown that the pairwise MRFs whose energy can be minimized in polynomial time using graph cuts correspond to a *submodular* energy function [87].

Unfortunately in computer vision applications of MRFs, it is often that non-submodular energy functions arise. In such cases, the minimization becomes NP-hard in general.

Boykov et al [18] introduce two kinds of large moves, the α-expansion and $\alpha\beta$-swap, for the optimization of MRF problems with multi-valued labels. They deal with a wide class of energies corresponding to MRFs with pairwise interactions. Ishikawa [66] extends the use of graph-cuts to the exact optimization of arbitrary convex pairwise MRFs. To accelerate graph-cuts in the case of dynamic MRFs (MRFs where the form of potential changes over time), the dynamic max-flow algorithm is proposed in [81, 82, 70]. Komodakis et al [92, 93] propose a primal-dual scheme to minimize the MRF energy, based on linear programming relaxation. Their approach has the advantage of computational efficiency.

Solutions for non-submodular energy functions can be inferred partially. [51] proposed roof duality to obtain a partial optimal labeling for quadratic pseudo-boolean functions. More recently, and based on the same concept, Quadratic Pseudo-Boolean Optimization (QPBO) was proposed [15, 158] and shown to be an effective approach. To deal with multi-label MRFs through QPBO, [80] proposed a method based on converting the multi-label MRF to an equivalent binary one. [108] proposed a method based on QPBO and move techniques, referred to as *fusion moves*. This technique is based on fusing two proposed solutions at each step of the algorithm, in order to achieve an energy

lower than that of either proposal.

Belief Propagation

Belief propagation (BP) was originally proposed in [131, 132]. It is a message-passing approach to inference on graphical models. It was originally used to perform exact MAP inference and/or max-marginal inference, on graphical models that can be represented by tree-structured factor graphs. In this case, belief propagation can perform inference in polynomial time.

Loopy belief propagation (LBP) is the application of belief propagation in general graphs [194, 190, 191, 30]. In general graphs, belief propagation has to be performed iteratively, and inference is approximate [190, 191]. LBP is not generally guaranteed to converge, but there exist conditions that guarantee convergence [118]. Despite the lack of guarantees of convergence, loopy belief propagation has performed well in a number of computer vision problems. Wainwright et al. [185] proposed a method inspired by the problem of maximizing a lower bound on the model energy. Based on this work, Kolmogorov [85] proposed a memory-efficient message-passing algorithm that achieves effective energy minimization in many practical applications.

Dual Methods

The problem of MAP inference in pairwise MRF models can be reformulated as an integer linear programming problem. Unfortunately this problem is NP-hard in general. However, several algorithms have been proposed for approximate MRF optimization, based on Linear Programming (LP) relaxations of such problems. It is generally infeasible to directly apply generic LP algorithms, such as interior point methods, to solve LP problems corresponding to MRF models in computer vision, due to the large number of variables. It is possible, however, to solve a dual to the original LP problem [95]. Methods that take such an approach include the message passing algorithm based on block coordinate descent proposed in [43], the Tree-Reweighted message passing (TRW) techniques [185, 85] and the dual decomposition (MRF-DD) approach proposed in [89, 91]. The tightening of the LP-relaxation has also been examined, in order to achieve a better optimum [43, 90, 96].

Inference on Higher-order MRFs

In recent years, a lot of research has concentrated in the case of MRFs with higher-order potentials. Such MRFs can better capture the statistics of several low-level vision problems [155], while they naturally arise to model other computer vision tasks. An example is the work presented later in this thesis, where a higher-order MRF naturally models the creation of cast shadows in an image. Inference on such MRF models remains however challenging.

One approach taken in order to minimize the energy of higher-order MRFs is to reduce the higher-order model to a pairwise one, by introducing extra variables. Inference in the pairwise model can be performed with one of the standard methods. This idea was first proposed in [154] (*variable substitution*), but the resulting pairwise models contained many non-submodular components, making inference difficult [15, 2]. An improved reduction method for second-order binary MRFs was proposed in [87] and [36] proposed an algebraic simplification of this approach. Based on the same concept, Ishikawa [65, 68] developed a technique that can reduce every higher-order MRF to a pairwise one. This technique can also deal with multi-label MRFs using fusion moves [67]. Approaches based on graph-cuts have also been proposed to tackle with more specific forms of higher-order MRF models [77, 78, 149].

Other authors have examined the use of belief propagation methods on higher-order MRF models. These techniques focus on taking advantage of specific classes of higher-order potentials to develop more efficient message-passing algorithms [99, 141, 175].

A third category of approaches proposed to deal with higher-order MRFs is based on the LP-relaxation formulation of the MRF energy minimization. In [88], the dual-decomposition framework [89] is applied to the case of higher-order MRFs. The result is a decomposition of the original MRF problem to a set of one master and several slave problems, with the master coordinating the solutions of the slaves. Inference algorithms are proposed to solve a class of MRF models with *pattern-based* potentials. Other recent approaches [156, 79] have also exploited the "sparseness" of certain forms of higher-order potentials.

Chapter 3

Extracting Shadows

In Chapter 2 we examined the literature in extracting the shadows from a single image. Despite the large number of approaches proposed, the problem is still challenging in the general case. The approach of intrinsic images does not separate the cast shadows from the shading component. The various illumination invariant representations do not generally offer satisfactory results in complex natural images; from our experiments, it seems that the best performing method illumination invariant representation is that of [33], which however does not perform well when camera calibration is not possible, and when the image color is distorted from factors such as JPEG compression. Some other methods need user-supplied hints which indicate where the shadows lie; the approach in [200] is overly complex and involves a learning stage, and the approach in [98] uses several assumptions that don't apply in arbitrary natural images.

In this chapter we will describe a simple new cue to aid shadow extraction (Sec.3.1.1), and we will describe a method to obtain an estimate of cast shadows that is sufficient for the task of illumination estimation. The goal of this method is not to provide a high-resolution estimate of shadow intensity (which could be used for shadow removal). Such an estimate could however be obtained from our results using a refinement stage, such as matting. The main advantage of the method proposed in this chapter is that it can be applied without assumptions about the camera or the lights illuminating the scene, it can be implemented efficiently and it does not depend on a training phase.

In the next section (section 3.1.1) we introduce a simple measure of brightness to aid in the extraction of shadows from the image. In section 3.1.2 we present our approach to obtain a set of confidence values that image segments

belong to cast shadows. Section 3.2 offers experimental evaluation of our approach and comparisons to current state-of-the-art methods, and section 3.3 concludes the chapter.

3.1 Our approach

We detect shadows by examining the change of image features across the borders of potential shadow regions. We start from the observation that light sources affect the whole image in a consistent way; therefore, edges due to cast shadows will generally exhibit characteristics that are consistent across the whole image, while edges due to other effects, such as albedo variations, will exhibit a more random behavior. To aid in the detection of shadows, we also utilize an appropriate measure of brightness, the *bright channel*. In the rest of this section, we explain our approach to shadow detection in more detail.

3.1.1 Bright Channel Cue

We first extract a measure of brightness from the image, the *bright channel* cue (defined similarly to the dark channel prior proposed in [57]):

$$I_{bright}(i) = max_{c\in\{r,g,b\}} \left(max_{j\in\Omega(i)}(I^c(j)) \right) \tag{3.1}$$

where $I^c(j)$ is the value of color channel c for pixel j and $\Omega(i)$ is a rectangular patch of size $m \times m$ pixels, centered at pixel i (in our experiments, $m = 6$.

The bright channel cue is based on the following intuition: The image values in patch $\Omega(i)$ are bounded by the incident radiance and modulated by the albedo at each pixel. However, in natural images, often a patch will contain some pixels with albedo that has high values in at least one color channel. By maximizing over color channels over all pixels in the patch, we reduce the effect of local variations of albedo within the image patch, getting a measure of brightness which is closer to the incident radiance at pixel i than the brightness at that pixel only.

We post-process the bright channel by choosing a white point I^{β}_{bright}, such that at least β % of the pixels are fully illuminated, corresponding to bright channel values of 1.0 (in our experiments, $\beta = 20\%$). Then the adjusted bright

channel values $\dot{I}_{bright}$ are:

$$\dot{I}_{bright}(i) = min\left\{\frac{I_{bright}(i)}{I^{\beta}_{bright}}, 1.0\right\} \tag{3.2}$$

Furthermore, the *max* operator in Eq. 3.1 implies a dilation operation, meaning that the dark regions in the bright channel image appear shrunk by $m/2$ pixels ($m \times m$ is the size of patches $\Omega(i)$). We correct this by expanding the dark regions in the bright channel image by $m/2$ pixels, using an *erosion* morphological operator [45]. An example of the bright channel is shown in Fig. 3.1.b.

3.1.2 Shadow detection

As mentioned above, we take advantage of the global nature of the effects of illumination to detect cast shadows. For example, if we examine features like the brightness ratio or the hue difference across the two sides of shadow edges, in a scene with a single light source we will notice that the values we observe are concentrated around a clearly defined center. Intuitively, the shadows are usually similarly dark and exhibit a similar color change everywhere when they are caused by the same light source. On the other hand, the same features across the sides of non-shadow edges are distributed in a much more random way in most images, because they are caused by albedo variations and other effects that are local in the image. The distribution of such features exhibits peaks that correspond to shadow borders in the image. Our goal is to detect such peaks.

All our computations to obtain confidence values for shadows are based on comparing image features on the two sides of potential shadow borders. To improve the robustness of such computations, when examining values on the two sides of pixel i lying on the border of segment S_j, we compare the *average* of values on two semi-circular patches P^i_{in} and P^i_{out} centered at pixel i, and oriented so that P^i_{in} is inside segment S_j and P^i_{out} is outside, as seen in Fig.3.1.d. We examine only border pixels where the ratio of the average bright channel value between the two patches P^i_{in} and P^i_{out} is larger than θ_e or smaller than $1/\theta_e$, to ignore pixels that do not correspond to image edges (in our experiments, $\theta_e = 1.2$).

We first obtain a segmentation $\mathcal{S}$ of the bright channel image $\dot{I}_{bright}$ [29]. From the set of segments in $\mathcal{S}$, we choose a subset of segments that are "good

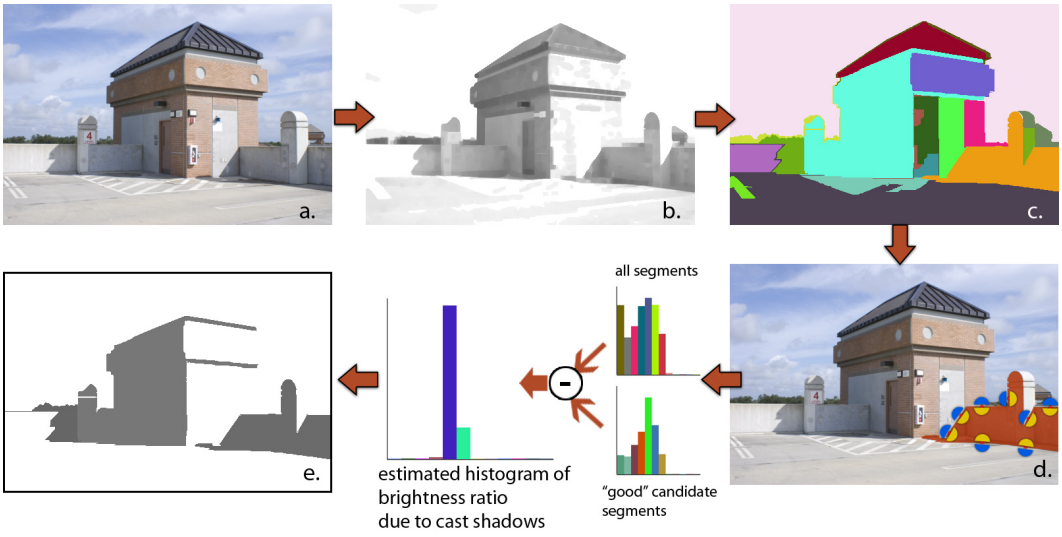

Figure 3.1: Shadow detection: a. original image (from [200]); b. bright channel; c. segmentation; d. for each segment border pixel, feature values are compared between two patches inside (yellow) and outside(blue) the segment; then we form histograms of the features observed for all segments, and for segments that are good candidates to correspond to shadows, and compute the difference of the two distributions; e. the final shadow estimate

candidates" to correspond to shadow regions. We define a "good candidate" for shadow as a segment where all three RGB color channels reduce in value across most of its edges, as we move from outside the segment towards the inside. We compute the confidence $q_{cand}(S_j)$ that a segment S_j is a "good candidate" to be a shadow as:

$$q_{cand}(S_j) = 1/|S_j| \sum_{i \in S_j} q(i; S_j), \tag{3.3}$$

where $q(i; S_j) = 1$ if the average of r, g and b color channels in P^i_{in} is darker than P^i_{out}, and 0 otherwise.

Let f be the chosen feature across segment borders (bright channel ratio or hue difference in our experiments) that depends on illumination. We create a histogram h^{all}_f of the values of feature f at all segment border pixels. We also create a histogram h^{good}_f of the values of feature f at each border pixel

i of each segment S_j, where each border pixel i contributes to the histogram proportionally to the confidence $q_{cand}(S_j)$. These two histograms represent the distribution of the values of feature f over all segment borders and over only segment borders that may be shadows. Normalizing them and taking their difference gives us a third histogram h_f^{diff} which corresponds to peaks in the distribution of feature f at borders in the set of "good candidates" that are not prominent in the distribution of f in the set of all segment borders. We expect that these peaks will correspond to the characteristics of the shadows: for example, if f is the bright channel ratio, then the peaks in h_f^{diff} will indicate how dark the shadows in the image are.

Based on the extracted histograms, we compute a confidence for each segment to correspond to a shadow. We approximate the distribution of feature f in h_f^{diff} by a mixture of normal distributions. Each component k of this mixture model is characterized by mean μ_k^f, variance σ_k^f and mixing factor π_k^f. We estimate these parameters through an Expectation-Maximization algorithm. To choose the number of distributions in the mixture we minimize a quasi-Akaike Information Criterion (QAIC). The confidence, based on a feature f, for segment $S_j \in \mathcal{S}$ is then defined as:

$$p^f(S_j) = \frac{1}{|\mathcal{B}_j|} \max_k \sum_{i \in \mathcal{B}_j} P_k \left(\Delta f(\mathcal{P}_1^i, \mathcal{P}_2^i) \right), \tag{3.4}$$

where $\mathcal{B}_j$ is the set of all border pixels of segment S_j, k identifies the mixture components, and, for patches $\mathcal{P}_1^i$ and $\mathcal{P}_2^i$ on the two sides of border pixel i, $P_k\left(\Delta f(\mathcal{P}_1^i, \mathcal{P}_2^i)\right)$ is the probability of observing the difference $\Delta f(\mathcal{P}_1^i, \mathcal{P}_2^i)$ in the average value of feature f between the two patches $\mathcal{P}_1^i$ and $\mathcal{P}_2^i$, according to mixture component k (and weighed by the mixture factor π_k).

If we know that there is only a single light source, as in the case of outdoor scenes, we can improve performance further by fitting a single normal distribution centered at the highest peak of h_f^{diff}.

The features used in our work are the bright channel value ratio and hue difference across patches $\mathcal{P}_1^i$ and $\mathcal{P}_2^i$. We compute the final confidence $p(S_j)$ that segment S_j is a shadow as:

$$p(S_j) = q_{cand}(S_j) \left(p^{bright}(S_j) + p^{hue}(S_j) \right) / 2. \tag{3.5}$$

The shadow intensity for a segment S_j is computed as the median of the bright channel value ratio of patch pairs inside and outside the segment

method	classification rate
our method (bright channel ratio)	87.7%
our method (hue)	86.7%
our method (combined)	**89.1%**
[200]	88.7%

Table 3.1: Pixel classification results with our method using different features, and with [200], on the UCF dataset ([200]).

(Fig.3.1.e).

This process is based on a segmentation of the image. In order to reduce our method's dependency on the quality of segmentations, we compute confidence values for different initial segmentations of the image. The final confidence value at pixel i is the mean of confidence values computed from each segmentation. Shadow detection can then be performed by thresholding the confidence value at each image pixel. In our experiments, we chose the threshold for shadow detection to maximize the classification rate on 100 training images from the UCF dataset [200].

3.2 Shadow Cue Evaluation

We evaluated our shadow detection approach quantitatively on the UCF dataset [200], which consists of 356 images and manually annotated ground truth for the cast shadows, using the same set of 123 test images as [200]. We also evaluated our approach on the 135 image dataset of [98]. In Fig.3.2 we show ROC curves with our method on both datasets and compare with [200], [31] and [98].

In figure 3.3 we show results using our shadow detection approach on the UCF dataset [200].

Our method performs similarly to [200] and significantly better than [31], which is affected by the low image quality and unknown camera sensors. One reason for the difference in performance to [98] is that the annotation of the ground truth in the dataset of [98] generally includes edges of cast but not attached shadows, whereas our method does not differentiate between the two. When the shadow is partially cast and partially attached, the ground truth in [98] contains only the partial boundary that corresponds to the cast

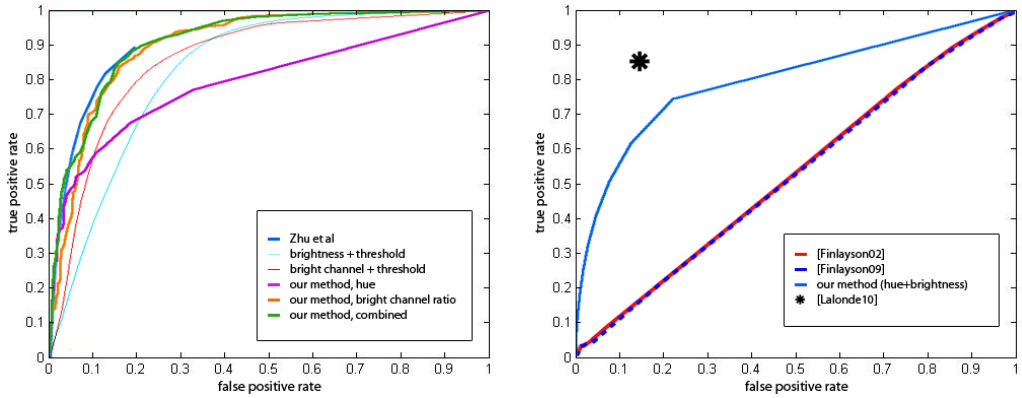

Figure 3.2: Comparison of our shadow detection method with different features and different methods ([200], [98], [31]). ROC curves computed on the dataset of [200] (top) and that of [98] (bottom).

shadow and thus cannot be matched correctly by our method that produces always closed shadow borders. In Table 3.1 we show pixel classification rates on the 123 test images from UCF dataset. To obtain these classification rates, we chose the decision threshold (see Sec.3.1.2) as the optimal threshold for a different set of 100 training images from the UCF dataset. The results show that our method is comparable to much more complex approaches. The average running time of our method for the test images in the UCF dataset is 2.7 sec which compares very favorably to the other methods.

The results in Fig.3.2 also justify our choice of the bright channel compared to simple image brightness (from the HSV color model), by examining the performance of each in shadow detection when used with simple thresholding.

3.3 Conclusions

In this chapter, we presented an approach to detect cast shadows in a single color image. Our approach does not make strong assumptions about the type of the scene (such as [98]) and is computationally efficient, taking a few seconds for each image. It also does not depend on training, such as [200]. The performance achieved is comparable to the aforementioned methods. In the next chapters, we will discuss illumination estimation utilizing the shadows extracted from the input image. The method described here is used in Chapter

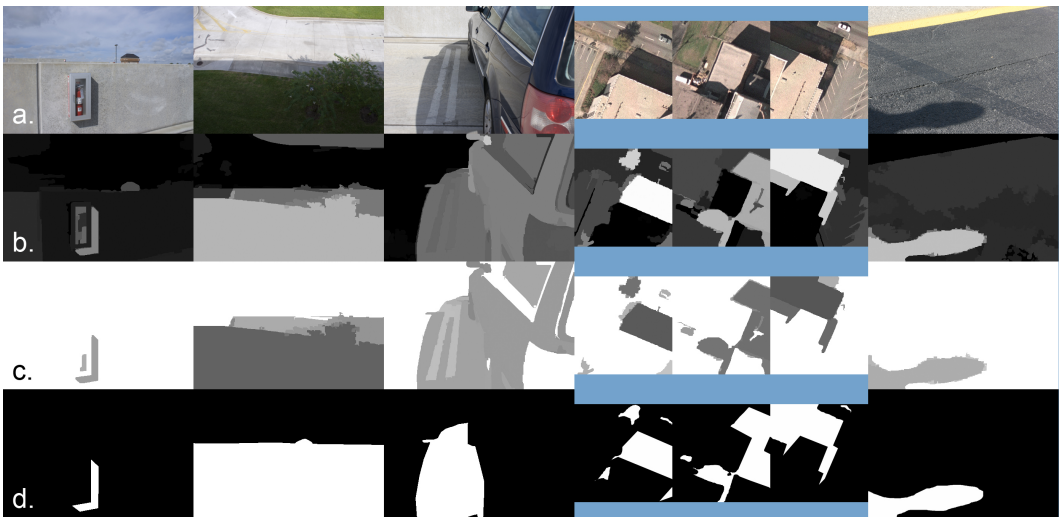

Figure 3.3: Shadow detection results: a) The original image, b) the confidence values we compute using our approach, c) the shadow values estimated by our approach after thresholding the computed confidence values, and d) the ground truth provided. Images are from the UCF dataset [200]. Notice that in some images, there are errors that are due to inaccuracies in the marking of ground truth (first image) or because of different treatment of attached shadows in our algorithm and in the ground truth (third image).

6 as input to the proposed illumination estimation method.

Chapter 4

Illumination Estimation through EM

In this chapter we present a new approach to illumination estimation. The base of this approach is the modeling of illumination as a mixture of probability distributions. Compared to previous approaches that use a set of samples of possible illumination directions ([166]) or a small set of point light sources, this allows us to better approximate the real illumination by modeling the approximate size of light sources, corresponding to the "softness" of shadow borders. At the same time, it enables us to perform illumination estimation robustly to large inaccuracies in knowledge of scene geometry, by an algorithm based on Expectation-Maximization (EM) [26].

We describe illumination as a mixture of von Mises-Fisher distributions. Our goal is to estimate the parameters of the distributions in this mixture. Given a single input image and a coarse model of the geometry of the scene, we first extract a set of illumination invariant features. Then illumination parameters are estimated using the EM algorithm. In the E-step, we first detect shadows, given the 2D cues (intensity variations and illumination invariant features), and input from the interaction of the light sources with the geometry. Afterwards, we update the expectations of the hidden variables that relate shadow pixels to the light sources in our model. Given these expectations, in the M-step we estimate the mean direction of the light source distributions, their intensities and shape parameters. Intensity and shape parameter estimation is performed using information directly from the image, in order to obtain an estimate more robust to the inaccuracies in geometry knowledge and detected shadows. The algorithm outputs a set of shadow labels and the

parameters that define the light source distribution.

Summarizing, the contributions of this algorithm are:

- we associate a mixture of von Mises-Fisher distributions with the generation of cast shadows in an image
- the above leads to a compact representation of the illumination which allows for robust estimation, relatively insensitive to inaccuracies in 3D geometry and shadow estimation.
- we integrate low-level cues obtained from illumination invariant features with 3D reasoning in a graphical model to enable shadow inference for textured surfaces

We validate our method by applying it to a dataset featuring images of simple objects in backgrounds that contain significant texture, under known and controlled illumination, as well as to a more challenging set of photographs of outdoor scenes involving geometrically complicated objects. We demonstrate that even when complex objects, such as a tree or a human, are modeled with simple bounding boxes, in natural scenes involving texture, our method is able to get a close approximation to the original illumination.

The remainder of this chapter is organized as follows: Sec.4.1 gives necessary background information; Sec.4.2 presents our model and the EM algorithm to perform inference with it; in Sec.4.4.1, shadow detection from illumination invariant features and their integration to our model is discussed; in Sec.4.5, the estimation of the parameters of the light source distributions in the mixture model is presented; results demonstrating the performance of our approach are presented in Sec.4.6, and in Sec.4.7 conclusions and future extensions are discussed.

4.1 Fundamentals

The inputs to our algorithm are the image I and a coarse 3D model of the geometry $\mathcal{G}$. We assume light sources are distant. Therefore, illumination can be approximated as a mixture of light distributions on a unit sphere of light directions. We model the light source distributions as von Mises-Fisher distributions. For each pixel i, a set R_i of N random 3D unit vectors expressing directions in 3D space is used to produce N samples of the illumination environment.

4.1.1 The von Mises-Fisher Distribution

A 3-dimensional unit random vector x (i.e., $x \in \mathbb{R}_3$ and $\|x\| = 1$) is said to have a 3-variate von Mises-Fisher (referred to as vMF henceforth) distribution [4] if its probability density function has the form:

$$f_v(x; \mu, \kappa) = \frac{\kappa}{4\pi \sinh k} e^{\kappa \mu^T x} \tag{4.1}$$

where μ is the mean direction, κ is the concentration parameter, $\|\mu\| = 1$ and $\kappa \geq 0$. The concentration parameter κ defines how strongly samples drawn from the distribution are concentrated around the mean direction μ. The von Mises-Fisher distribution is the equivalent of a Gaussian distribution on a sphere, and it is used widely in directional statistics.

4.2 Model Description

In this section, we formulate the generation of cast shadows as a mixture of vMF distributions on the unit sphere, and present the general EM framework to estimate the parameters of this mixture model.

We assume that light sources are distant. Let i be a pixel of the original image. We sample the incoming radiance at this pixel along N randomly chosen directions. The image value $I(i)$ at pixel i can be discreetly approximated by the sum of the contributions of the light sources along each sampling direction $\mathbf{r}_k$.

A light source contributes to the incoming radiance along direction $\mathbf{r}_k$ only if the ray from the 3D position of pixel i along direction $\mathbf{r}_k$ does not intersect the geometry of the scene $\mathcal{G}$. We repeat from Eq.2.3 the definition of the visibility term $V_{\mathbf{p}}(\mathbf{r}_k)$ for a ray along direction $\mathbf{r}_k$, originating at the 3D point $\mathbf{p}$ that projects to pixel i:

$$V_{\mathbf{p}}(\mathbf{r}_k) = \begin{cases} 0, & \text{if ray from } i \text{ along } \mathbf{r}_k \text{ intersects } \mathcal{G} \\ 1, & \text{otherwise} \end{cases} \tag{4.2}$$

Assuming Lambertian reflectance, the image intensity $I(i)$ at pixel i could ideally be synthesized from the sum $\hat{I}(i)$ of the contributions of the light source

distributions:

$$\hat{I}(i) = \sum_{k=1}^{N} \left(V_{\mathbf{p}}(\mathbf{r}_k) \max\{\mathbf{n}_i \cdot \mathbf{r}_k, 0\} \sum_{j=1}^{M} l_j(\mathbf{r}_k) \right), \tag{4.3}$$

where M is the number of distributions used to approximate the illumination, $\mathbf{n}_i$ is the normal corresponding to pixel i and $l_j(\mathbf{r_k})$ is the contribution of light distribution j along direction $\mathbf{r_k}$. We model each light distribution j as a von Mises-Fisher distribution, with mean direction μ_j, concentration parameter κ_j and intensity (mixing factor) α_j, and therefore:

$$l_j(\mathbf{r}_k) = \alpha_j f_v(\mathbf{r}_k; \mu_j, \kappa_j). \tag{4.4}$$

We assume that $\sum_{j=1}^{M} \alpha_j = 1$. Therefore, we can describe the illumination environment using the set of parameters $\theta = \{\mu_1, \kappa_1, \alpha_1, ..., \mu_M, \kappa_M, \alpha_M\}$, which we need to estimate.

4.3 The EM Algorithm

The problem of estimating the illumination from cast shadows, given the above model, can be regarded as the problem of estimating the mixture of vMF distributions defined in Eq.4.3. For this estimation problem we use the EM algorithm, which has been used widely to estimate the parameters of mixture models due to its simplicity and numerical stability. Our formulation closely resembles the soft-assignment scheme described by Banerjee et al. [4] to estimate the parameters of a mixture of vMF distributions.

Let $X = \{x_1, ..., x_P\}$ be the set of pixels in the image and $L = \{L_1, ..., L_M\}$ the set of light source distributions. For each pixel, a set $R = \{\mathbf{r}_1, ..., \mathbf{r}_N\}$ of N sampling directions is used. The sampling directions $\mathbf{r}_i$ are chosen randomly. Our algorithm initializes the cluster means μ_j randomly for each light source distribution k, the concetration parameters to $\kappa_j = 1$ and the intensities to $\alpha_j = \frac{1}{|L|}$. Then the following steps are repeated:

E-step

At each iteration, in the E-step we detect shadow pixels, calculating the probability $P(s_i|I, \theta)$ that pixel x_i is in shadow, given the current estimate of the

parameters θ, and then we estimate the new values of the parameters for only one distribution j in each iteration.

The probability that pixel i is in shadow cast due to light distribution j, given our current estimate of the parameters θ, expresses the probability that pixel i has been labeled as shadow, and that the expected shadow intensity by all distributions other than j does not explain pixel i:

$$q_j(x_i; I_S, \theta) \leftarrow P(s_i|I_S, \theta) \max\left\{ I_S(x_i) - \sum_{k=1, k\neq j}^{M} \hat{I}^k(x_i; \theta), 0 \right\}, \tag{4.5}$$

where $I_S(x_i)$ is the shadow intensity value, as estimated in Sec.4.5 (for the first EM iteration, when there is no such estimate, $I_S(x_i) = 1$ if pixel i is labeled as shadow, and 0 otherwise). $\hat{I}^k$ are the image intensities as synthesized by rendering the scene using the k-th light source distribution.

To synthesize $\hat{I}^k$, we sample light source distribution k using the accept-reject algorithm to generate a set of incoming light directions $R_k = \{\mathbf{r}_1^k, ..., \mathbf{r}_Q^k\}$ for each pixel i, where Q the number of samples. Then, the synthesized image intensity for pixel i is given by:

$$\hat{I}^k(i; \theta) = \sum_{r \in R^k} V(\mathbf{r}^k) \max\left\{ \mathbf{r}^k \cdot \mathbf{n}_i, 0 \right\}, \tag{4.6}$$

which is another way to look at the generation of shadows in Eq.4.3.

We update the expectation for each hidden variable $h_{i,k}$, associated to sample direction $\mathbf{r}_k$ for pixel i, using the following rule:

$$p_j(h_{i,k}; I, \theta) \leftarrow q_j(x_i; I, \theta) \frac{1 - V_{\mathbf{p}}(\mathbf{r_k})}{\sum_{m \neq k} (1 - V_{\mathbf{p}}(\mathbf{r_m}))} \times \frac{f_v(\mathbf{r_k}; \mu_j, \kappa_j)}{\sum_{n=1}^{M} f_v(\mathbf{r_k}; \mu_n, \kappa_n)}. \tag{4.7}$$

M-step

In the M-step, we update the parameters θ for each k=1...M. The mean directions μ_j are estimated by:

$$\mu_j = \frac{1}{P} \sum_{i=1}^{P} \left(\frac{1}{|R_i|} \sum_{k \in R_i} p_j(h_{i,k}; I, \theta) \mathbf{r_k} \right) \tag{4.8}$$

The concentration parameters κ_j and the intensities α_j are estimated directly from the image, as described in Sec.4.5.

4.4 Shadow Detection in the E-step

In the E-step of the EM algorithm described above, we used the probability that a pixel i corresponds to a cast shadow. We formulate the estimation of these probabilities as a labeling problem, where shadows are identified by assigning a set of binary labels $S = \{s_i | i \in P\}$ to image pixels P:

$$s(i) = \begin{cases} +1, & \text{if pixel } i \text{ is in shadow} \\ -1, & \text{otherwise} \end{cases} \tag{4.9}$$

As mentioned, we use a number of 2D cues coming from illumination invariant representations of the image (fig.4.1), combined with information from 3D reasoning to estimate the shadow labels. The probability of the shadow labels S is modeled as:

$$P(S|I,\theta) \propto \prod_i \left(P(s_i|\theta) \prod_{j \in \mathcal{N}(i)} P(s_i, s_j|I) \right), \tag{4.10}$$

where $\mathcal{N}(i)$ are the pixels neighboring to i. The term $P(s_i|\theta)$ models the probability that pixel i is in shadow given the current estimate of the illumination, and enforces geometrically meaningful shadows. It is approximated using the expected shadow values $\sum_{j=1}^{M} I_S^j$, spatially smoothed. The term $P(s_i, s_j|I)$ represents the joint probability of labels s_i and s_j for neighboring pixels i and j, given some image features for these pixels. This term encodes our estimate that the corresponding image gradient should be attributed either to shadow or to texture. The computation of shadow borders is discussed in section 4.4.1.

The distribution of shadow labels can be represented by a factor graph which corresponds to a 2D lattice. Inference to find the labels s_i at each step is performed using loopy belief propagation.

4.4.1 Identifying Shadow Borders

In order to identify shadow borders, as required for our shadow detection approach, we utilize illumination-invariant representations of the original image.

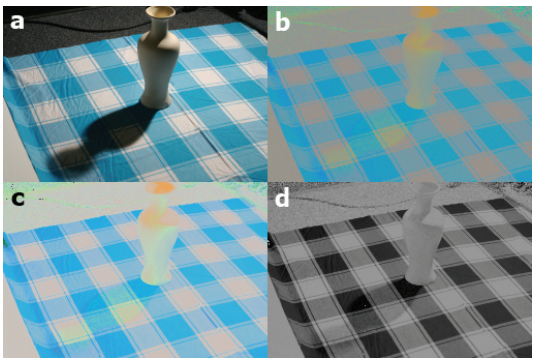

Figure 4.1: Illumination invariant images: a) original image, b) normalized rgb, c) $c_1c_2c_3$, d) the 1d illumination invariant image obtained using the approach in [33]. Notice that in all three illumination invariant images, the shadow is much less visible than in the original.

Illumination-invariant representations were discussed earlier in Section 2.2.1. The ones we choose to use here are the Normalized RGB and $c_1c_2c_3$ representations and the illumination invariant described in [33].

A border which appears in the original image but not in the illumination invariant images is a border which can be attributed to illumination effects. Therefore, to identify potential shadow borders, edges are detected in the original image and each edge is checked against the illumination invariant images. Calculating edges as simple finite difference approximations to gradients leads to a lot of noise, detecting edges that are not important. To solve this, we apply a smoothing filter to the original image, and then use the Canny edge detector to perform edge detection.

We do not calculate similar edge maps from the illumination invariant images. Instead, for each pixel that lies on an edge in the original image, we compare the difference of the average values of the illumination invariants along the direction of the gradient in the original image. Thus the shadow border map is defined as:

$$e_s(x,y) = \begin{cases} 1, & \text{if } \|\nabla I\| > \tau_0 \text{ and } |\Delta I^{(k)}_{invar}| < \tau_k \\ 0, & \text{otherwise} \end{cases} \tag{4.11}$$

where $\Delta I^{(k)}_{invar}$ is the result of a step filter oriented along the image gradient and applied to illumination invariant image k, $k = 1, 2, 3$. The parameters $\tau_0, ..., \tau_3$

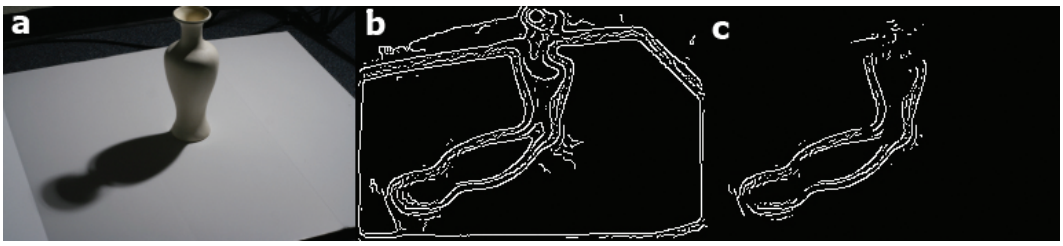

Figure 4.2: Shadow borders: a) original image, b) estimate using only 2D cues, c) refined estimate after first iteration

are learned directly from data, as the values that best separate shadow borders from edges not related to shadows in the training set. We prefer this method over directly comparing with edges in the illumination invariant image (as in [33] for example) in order to deal with very soft shadows and edge localization differences in the original and the invariant image.

Because the illumination invariant features often either contain some illumination information, or omit some information that is not related to illumination, the shadow borders detected using the above method generally include borders that are not related to shadows (figure 4.2.b). To alleviate this problem, we take advantage of the current estimate of the illumination to remove unreasonable shadow borders, by defining the final shadow edges as:

$$E_s(x, y) = e_s(x, y)\|\nabla I_S\| \tag{4.12}$$

where I_S is the shadow map expected from our current estimate of the illumination parameters, θ and the rough geometry G, smoothed with a gaussian filter. The refined shadow borders after the first iteration are shown in figure 4.2.c.

4.4.2 Integrating Shadow Borders to our Model

Shadow borders are integrated in our model by the term $P(s_i, s_j|I)$ in Eq.4.10, which defines the probability of the pair of labels for pixels i and j given the corresponding image features. If pixels i and j do not belong to an image border, then this term enforces uniformity of labels, so it becomes:

$$P_{uniform}(s_i, s_j|I) = \begin{cases} 1 - \theta_1, & \text{if } s_i = s_j \\ \theta_1, & \text{otherwise} \end{cases} \tag{4.13}$$

If one of i and j belongs to a shadow edge, this term enforces a transition in the labels from i to j. The probability of the pair of labels of i and j becomes:

$$P_{border}(s_i, s_j|I) = \begin{cases} 1-\theta_2, & s_i = +1,\ \|I_i\| < \|I_j\| \\ 1-\theta_2, & s_j = +1,\ \|I_j\| < \|I_i\| \\ \theta_2, & \text{otherwise} \end{cases} \tag{4.14}$$

In the above equations, $s_i = +1$ if pixel i is in shadow and the constants θ_1 and θ_2 are learned from the training data. We do not assume that $P(s_i, s_j|I)$ has a distribution dependent on the difference of the intensities of i and j in order to make possible the detection of dim shadows. Often the intensity changes over falsely detected shadow edges are much larger than the ones over real shadow borders.

4.5 Estimating κ and Intensity

Estimating the concentration parameter κ for a mixture of vMF distributions requires significant approximations [4]. In our model, it becomes even more difficult because the values of the samples are not individually observed; instead, only their per-pixel sums are known. It is easy to observe, though, that there is a clear connection between the shadow edge gradients, as they appear in the image, and the concentration parameter of the light source distributions. We exploit this connection to derive an estimator for κ.

4.5.1 κ Estimation

Let I_S^k be the image of the shadow intensities attributed to light source distribution k. Let i and j be two neighboring image pixels and $\Delta I_S^k(i,j) = I_S^k(i) - I_S^k(j)$ the finite approximation to the shadow intensity gradient between pixels i and j because of light source k. Using a linear approximation for e^x, we derive from Eq.4.3 the following relation connecting $\Delta I_S^k(i,j)$ and the parameter κ_k:

$$\kappa_k \geq \frac{|\Delta I_S^k(i,j)| - \left(\sum_{\mathbf{r}\in R_1} V_{\mathbf{p}}(\mathbf{r}) - \sum_{\mathbf{r}\in R_2} V_{\mathbf{p}}(\mathbf{r})\right)}{\sum_{\mathbf{r}\in R_1} V_{\mathbf{p}}(\mathbf{r}) \max\{\mathbf{r}\cdot\mu_i, 0\} - \sum_{\mathbf{r}\in R_2} V_{\mathbf{p}}(\mathbf{r}) \max\{\mathbf{r}\cdot\mu_i, 0\}} \tag{4.15}$$

where R_1 and R_2 are the sampling directions for illumination at pixels i and j respectively.

We evaluate Eq.4.15 only for the neighboring pixel pairs i and j for which the absolute value of the denominator:

$$\left| \sum_{\mathbf{r} \in R_1} V_{\mathbf{p}}(\mathbf{r}) \max\{\mathbf{r} \cdot \mu_i, 0\} - \sum_{\mathbf{r} \in R_2} V_{\mathbf{p}}(\mathbf{r}) \max\{\mathbf{r} \cdot \mu_i, 0\} \right| \geq \epsilon, \tag{4.16}$$

for some very small ϵ. Given the above, this means that we select only the pixel pairs were shadow variations are expected given the light direction μ_k and geometry $\mathcal{G}$. These are the areas that are actually informative about the value of κ.

To estimate the true shadow image gradient ΔI_S^k due to light source distribution k, for a shadow edge pixel i, we project the observed gradient $\nabla I(i)$ along the direction of the synthesized shadow gradient $\nabla \hat{I}^k$. As mentioned earlier, $\hat{I}^k$ is the shadow synthesized from the geometry and light source distribution k.

$$\Delta I_S^k(i) = \frac{\nabla I(i) \cdot \nabla \hat{I}^k(i)}{\alpha_k \|\nabla \hat{I}^k(i)\|}, \tag{4.17}$$

where α_k is the current estimate of the intensity of light source k. $\hat{I}^k$ has been smoothed using a gaussian kernel.

To estimate κ, we compute the estimate from Eq.4.15 for all pixels located around the identified shadow borders. The estimate of κ_k is selected to be the maximum of all per-pixel estimates from Eq.4.15. In practice, we discard the top 1% of values as outliers and select the maximum of the remaining values as the value of κ_k.

4.5.2 Light Intensity Estimation

Intensity estimation for light source k is also based on shadow borders. For each pixel i with coordinates (x, y) that lies on an identified shadow edge, $\nabla \hat{I}^k(x, y)$ defines a direction perpendicular to the *synthesized* shadow edge. We select two samples $\mathbf{p}_1 = [x, y]^T + t_1 \nabla \hat{I}^k(x, y)$ and $\mathbf{p}_2 = [x, y]^T - t_2 \nabla \hat{I}^k(x, y)$. Starting with $t_1 = t_2 = 0$, we individually increment t_1 and t_2 until we find a minimum of $\nabla \hat{I}^k(\mathbf{p}_1)$ and $\nabla \hat{I}^k(\mathbf{p}_2)$ respectively. Then we assume that $\mathbf{p}_1$ lies inside the shadow umbra and $\mathbf{p}_2$ is outside the shadow. The intensity difference $\Delta I(i) = I(\mathbf{p}_2) - I(\mathbf{p}_1)$ is an estimate of the shadow intensity. The

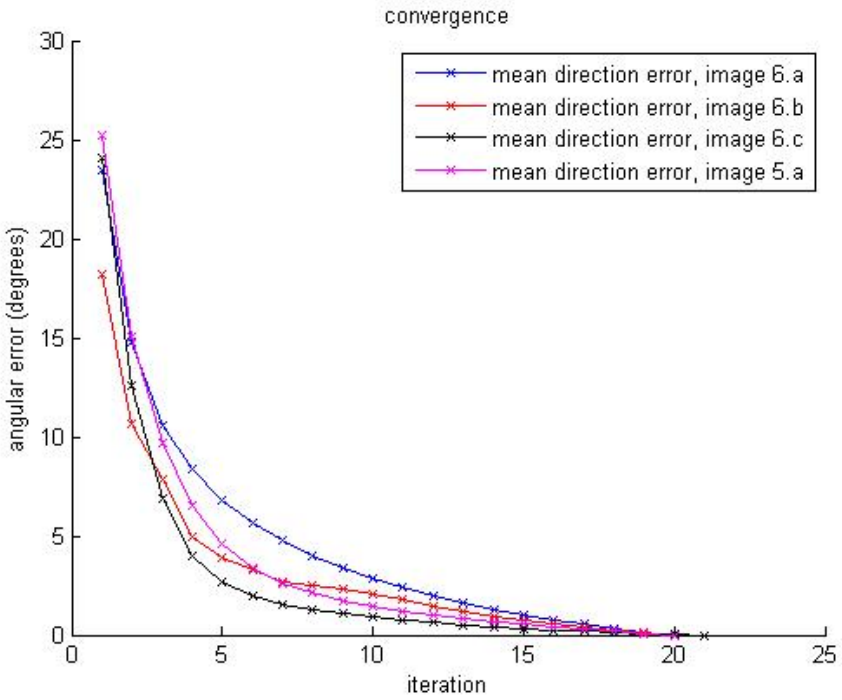

Figure 4.3: Convergence: the plot shows the mean difference (in degrees) between the estimated light source directions for each iteration and the final parameter values from our algorithm

light source intensity α_k is set to the mean value of $\Delta I(i)$ over all shadow edge pixels. Intensities are normalized so that $\sum_k \alpha_k = 1$.

4.6 Results

For all of our experiments, 200 random samples of the illumination sphere per pixel were used. A maximum of 40 EM iterations and 1500 iterations for the belief propagation in the factor graph were performed. The average running time of the algorithm was 3-5 minutes per image (For performance reasons, several EM iterations were performed before successive applications of belief propagation in the E-step). On average, our algorithm needed 15 to 20 EM iterations per light source distribution to converge (see Fig.4.3). The

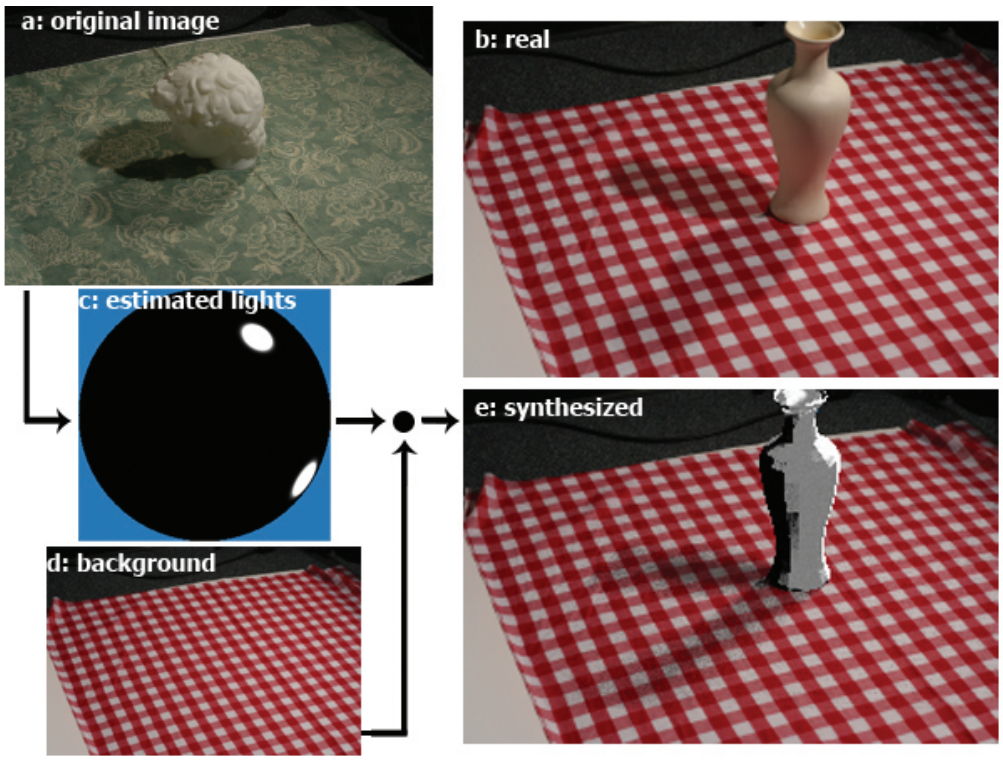

Figure 4.4: Comparison of real and synthesized shadows: a,b) photographs under same illumination, c) estimated illumination from (a), d) a picture of the background with the object removed, e) a 3D model of the original object rendered with the estimated illumination, and superimposed on the background image of (d). The shadows in this image are rendered with the estimated illumination and cast on the background image.

3D models we used to approximate the geometry consisted of 8 to 15 polygons each for all results presented here.

A dataset of 58 pictures captured in a controlled environment, using various background textures, was used to evaluate the algorithm. The geometry of the objects and the illumination environment were both known in these cases. 5 of the images were used to learn the parameters for shadow border detection and the rest were used for testing. Results in some representative examples of images are displayed in Fig.4.5 and 4.6. In Fig.4.4 we show the augmentation of a real scene with a synthetic object, compared to the image of the actual

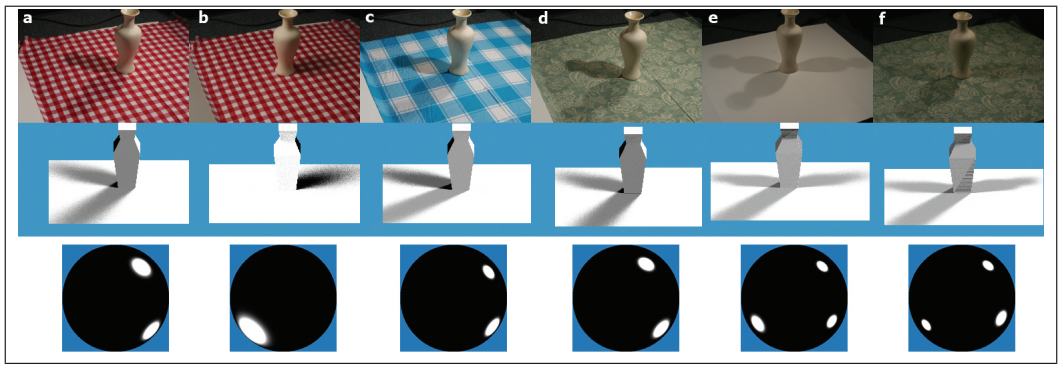

Figure 4.5: Results for shadows cast on different textures with our method: the original images are in the first row, the shadows as rendered from the coarse 3D model (used for the estimation) and the estimated illumination, using the same viewpoint as the original image, are in the second row, and the third row shows the illumination sphere as viewed from the top of the scene. Images a, c and d have been captured using the same lights setup. The mean difference of light source directions from these 3 images and 2 more with the same original illumination and different background (not shown here) was 4.92 degrees. Images e and f were captured using 3 light sources.

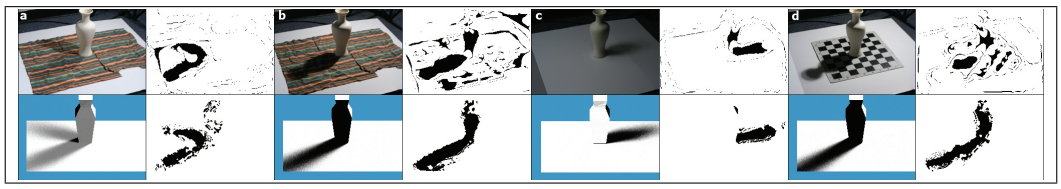

Figure 4.6: Results with our dataset: For each of 4 input images, clockwise, the original image, the labeling before the first iteration (using only 2D cues), the final labeling, and the coarse 3D model rendered with the estimated illumination are displayed. Notice that even in a difficult case, such as image d, where the initial shadow labeling is very poor, our algorithm is able to discover the shadows and estimate the illumination.

object in the scene.

The algorithm was also tested against 3 images of natural scenes. The parameters used were the same ones used with our collected dataset. These images were taken outdoors, so they involved only one major light source, the sun. However, they also involved texture, complex backgrounds and very complicated geometry, which we approximated with simple box-like models. The results are shown in Fig.4.7.

The mean direction of the light source distributions is estimated accurately from shadows cast on surfaces with a variety of textures. The mean error for directions estimated under the same illumination, for the same object but with 5 different textured backgrounds (three examples are in Fig.4.6 a, c and d) was 4.92 degrees. The estimation of the concentration parameter κ is often inaccurate, especially in the presence of texture. A better separation of texture and shadow is required for better estimation of κ.

The number of distributions used in the mixture model does not affect the results substantially. If the number of distributions used is larger than necessary, the distribution means tend to cluster together in clusters that correspond to the actual lights. When the number of distributions is less than that of the major light sources, our model tends to select some of the shadows, leaving others unexplained.

4.7 Conclusions

In this chapter we described a new method to identify cast shadows and model their generation using a mixture of vMF distributions. Our model requires a single input image and a coarse 3D model to describe the scene geometry, and is robust to poor geometry information and poor initial shadow labeling. Furthermore, the illumination estimation results are stable regardless of the texture of the surfaces on which shadows are cast. The ability to model scene geometry with 3D models as coarse as simple bounding boxes would make it possible to use our algorithm in combination with e.g. object detectors instead of full geometry, combined with a camera position estimate. Our results show that our method can be useful not only in estimating illumination for augmenting a real scene with synthetic objects, but also for tasks such as segmentation and more general reasoning about the 3D scene represented by an image. Interesting extensions to this work would include extending the EM method to handle complex natural illumination (which would mainly require

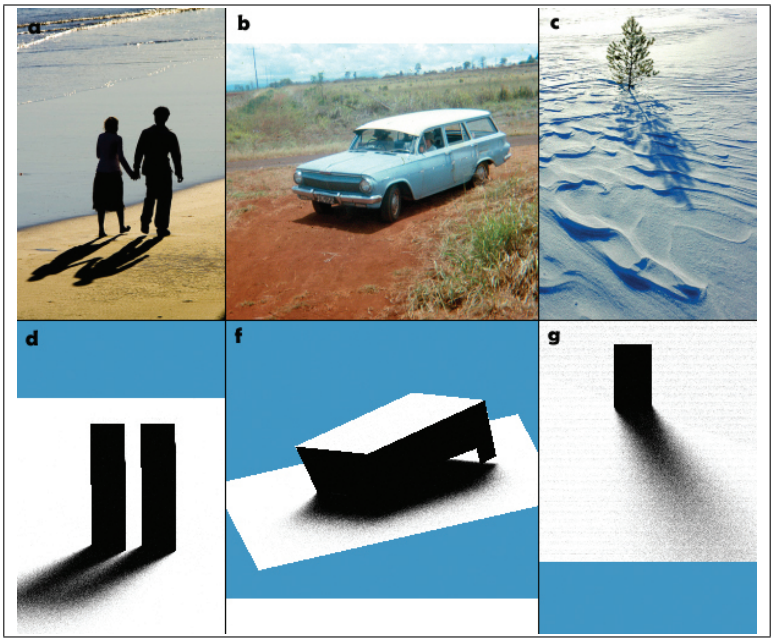

Figure 4.7: Results with photographs from Flickr. Top: original image, Bottom: the rough 3D model used for illumination estimation, rendered with the estimated illumination under the same viewpoint. Notice that despite the inaccuracies of the 3D models (mainly boxes), the shadows match well.

changes in the M-step estimators), applying the EM method with the bright channel shadow cue to improve results, and examining a tighter integration of shadow estimation and the feedback from 3D reasoning based on illumination. Such a tighter integration, under a different model for illumination, is presented in the Chapter 6, in a model that has the ability to easily incorporate different facets of the illumination estimation problem.

Chapter 5

Illumination from Shadow Edges

5.1 Introduction

In this chapter we discuss illumination estimation in general scenes and associate it with the existence of shadow edges. Most general illumination estimation methods from shadows (or shading) associate a parametrization of illumination with the per pixel intensity of shadows or shading [166]. As a result, estimating illumination from shadows in a general scene generally needs a way to separate shadows from scene albedo and other effects as initial input. Two significant types of errors can be introduced this way: errors in the initial shadow estimate propagate throughout the illumination estimation process, altering the final results, even in the case where the shadow estimate is refined during illumination estimation [129]; on the other hand, the knowledge of scene structure may not be adequate to explain a lot of correctly detected shadows in complex scenes, leading to erroneous illumination solutions that try to explain every observed shadow with inadequate geometry data. In this chapter we propose a way to couple shadow and illumination estimation, trying to detect only the shadow edges that are relevant to the provided geometry, as part of the illumination estimation process. This leads to an illumination estimation algorithm that performs on par with or better than the state of the art, even when scene geometry knowledge is limited, while being less dependent on obtaining an initial shadow estimate (we can even avoid obtaining an estimate of shadow edges altogether, as we show in the results in this chapter), and having lower computational complexity than state-of-the-art methods. In this approach, illumination estimation is posed as the minimization of an energy

function, and coupled with the detection of salient shadow edges.

We define this energy function to correspond to the quality of the matching between the observed shadow edges in the image and the shadow edges expected by the illumination solution. We extract the set of observed shadow edges by comparing gradients in the original image and two illumination invariant representations of it; in the limit, our approach can work without performing any shadow edge detection at all, assuming that all image edges are potential shadow edges (as we demonstrate in figure 5.7). The potential shadow edges are encoded in a shadow edge confidence map, and a simple approach is described to minimize the solution energy given this map, obtaining the illumination parameters that correspond to a good matching of the expected shadow edges with observed image edges.

The contributions of this work are the following:

- We explicitly associate illumination with shadow edges instead of per-pixel shadow intensities. This allows our approach to ignore errors in shadow detection, and concentrate only on potential shadow silhouettes that are meaningful given the scene geometry.

- This fact further allows our approach to estimate illumination using 3D geometry that only partially models a complex scene; for example, approximate knowledge of a single shadow-casting object and the rough shape of the surface its shadow is cast on can be adequate to estimate illumination in a larger, complex scene.

- Our approach is robust to inaccurate knowledge of 3D geometry, allowing us to model objects in real images using very coarse geometry, such as 3D bounding boxes. Our quantitative results demonstrate the robustness of our method with regard to geometry inaccuracies.

We present both quantitative and qualitative results. Quantitative results show the accuracy of our approach when estimating illumination in a synthetic dataset. Qualitative results show how our method performs in a set of real images collected from Flickr, and are compared to the results obtained by [129] (which we describe in the next chapter). At the same time, the computational cost of our approach is significantly lower than comparable approaches, corresponding to a much simpler implementation.

This chapter is organized as follows: in section 5.2 we formulate illumination estimation as the minimization of an energy function that measures

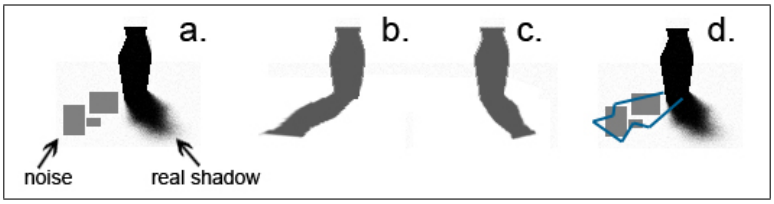

Figure 5.1: Motivation for using shadow edges instead of pixel intensities: We examine the example input image in (a) which includes a well-defined shadow and some noise. We examine the error of two possible light configurations that produce the images in (b) and (c). If the current estimate of ambient illumination is high, leading to dim gray shadows, then computing the per-pixel error between the expected shadow and the observation would produce a lower error for (b) than for (c). This is because most pixels in (b) has similar values as the corresponding pixels in the observed image, with a few white pixels in between. On the other hand, most shadow pixels have significantly different values than those produced by the configuration of (c). However, the shadow outline produced by the configuration of (b) is different than the observed image edges, while (c) matches the image edges well. Therefore, penalizing differences in the predicted and observed orientation of edges will favor the correct configuration (c). The shadow edges effectively lead to an energy with less local minima than a similar formulation using pixel intensities.

the quality of the match between the expected shadow gradient and an edge map extracted from the image. Section 5.3 describes how we obtain this edge map, while in section 5.4 we describe energy minimization and we extend our solution to the case of multiple light sources. Results are presented in section 5.5, while section 5.6 concludes the chapter.

5.2 Formulation

We will first examine the case where the scene is illuminated by a single distant light source, with direction $\mathbf{d}_0$ and intensity α_0. Let $\mathcal{E} = \{e_i\}$ be a set of edges detected from the original image, and $Q(e_i) \in [0, 1]$ be a confidence value that edge e_i is generated by a cast shadow (larger values indicate higher confidence). A geometric model $\mathcal{G}$ is also known. Geometry $\mathcal{G}$ may model only a small part of the scene and may be approximate - e.g. in many of our experiments we approximate objects by 3D bounding boxes.

Our goal is to find the light parameters $\theta_{\mathcal{L}} = (\mathbf{d}_0, \alpha_0)$ that produce a

shadow with shadow borders $\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})$ that:

- best coincide with image edges that have high confidence values $Q(e_i)$ to belong to shadows, and
- have a similar orientation with the corresponding observed image edges.

We express this requirement by defining an energy for each set of light parameters:

$$E_{match}(\theta_{\mathcal{L}}) = \frac{1}{\left|\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})\right|}\left(\sum_{i\in\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})}(1-Q(i)) + \sum_{i\in\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})}\widehat{\nabla I_i, \mathbf{e}_i}^2\right), \qquad (5.1)$$

where $\widehat{\nabla I_i, \mathbf{e}_i}$ is the angle between the observed image gradient ∇I_i at pixel i and the direction of the synthetic shadow edge at i, $\mathbf{e}_i$.

Notice that in this formulation, we have already removed several important requirements of traditional illumination estimation methods:

1. We do not need to know or estimate the intensity of ambient illumination
2. We are not defining the energy over all possible shadow edges in the scene, but only for that set of edges that is generated by the geometry $\mathcal{G}$ and the set of light parameters $\theta_{\mathcal{L}}$.
3. We do not need to estimate the intensity of light sources while estimating light source directions, because the set of edges $\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})$ depends only on the light source direction. The light source intensity can be included in the energy minimization (see Eq.5.2) or, as we preferred here, it can be estimated after the light source directions have been estimated.

Therefore, the matching cost $E_{match}(\theta_{\mathcal{L}})$ only depends on the light directions and the confidences assigned to observed shadow edges.

If we wish to estimate light source intensity α_0 concurrently with light source direction, we can minimize the sum of $E_{match}(\theta_{\mathcal{L}})$ and a term $E_{\alpha}(\theta_{\mathcal{L}})$:

$$E_{\alpha}(\theta_{\mathcal{L}}) = \sum_{i\in\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})}\left(\alpha_0 - \frac{\overline{I}_{out} - \overline{I}_{int}}{\max\{\mathbf{n}_i\cdot\mathbf{d}_0, 0\}}\right)^2, \qquad (5.2)$$

where $\overline{I}_{in}$ and $\overline{I}_{out}$ are the mean pixel intensities of two image patches placed on the two sides of pixel i, in the inside and outside of the expected shadow

respectively, $\mathbf{n}_i$ is the normal vector at pixel i, as given by the provided 3D geometry (if any for that pixel), and $\mathbf{d}_0$ is the light source direction. Eq.5.2 assumes Lambertian reflectance - our estimates, however, do not deteriorate significantly when this assumption is violated, and no such assumption is necessary to estimate only the light source directions.

We therefore manage to associate the light source directions with only the subset of observed edges in the image that matches the shadow borders $\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})$ produced by the current illumination estimate. One obvious issue that arises, though, is that in some cases there are trivial solutions that do not produce any cast shadows and such solutions will be preferred because they lead to minima of $E_{match}(\theta_{\mathcal{L}})$. To avoid this, we encourage solutions that have a larger number of well-explained shadow edge pixels, by defining the final energy to be minimized as:

$$E(\theta_{\mathcal{L}}) = \frac{1}{1+|\hat{\mathcal{E}}_g|}\left(E_{match}(\theta_{\mathcal{L}}) + w_\alpha E_\alpha(\theta_{\mathcal{L}})\right), \tag{5.3}$$

where w_α is a weight and the term $w_\alpha E_\alpha(\theta_{\mathcal{L}})$ can be omitted if there is no need to estimate light directions and intensities *concurrently*. The set $\hat{\mathcal{E}}_g$ is the set of all the expected shadow edge pixels in $\hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})$ that coincide with observed edges of high confidence:

$$\hat{\mathcal{E}}_g = \left\{ e_i \in \hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G}) \;\mid\; Q(e_i) > \theta_Q \right\}, \tag{5.4}$$

where θ_Q is a confidence threshold. We set $\theta_Q = 0.5$ in our experiments.

5.3 Extracting the edge map

The main term of the energy we want to minimize is $E_{match}(\theta_{\mathcal{L}})$, which is mainly a sum of confidence values along the expected shadow borders, the form of the confidence map is important for finding the light parameters that minimize $E(\theta_{\mathcal{L}})$.

Let $\mathcal{Q} = \{Q(i)\}$ be the confidence map. For each image pixel i, the confidence $Q(i)$ expresses the probability that i belongs to a shadow edge, if there is an edge at i, or the probability that a shadow edge lies in the vicinity of i. Map $\mathcal{Q}$ must contain confidence values that smoothly increase as we approach observed edges in the image, in order to allow effective minimization of $E(\theta_{\mathcal{L}})$.

Figure 5.2: Extracting the edge map: a.The original image; b. The edge confidence map $\mathcal{Q}$ (right) extracted from the original image. Brighter pixels indicate higher confidence; c. The gradient directions after smoothing (used to penalize the expected shadow gradient). The x and y components of the gradient are encoded in the red and green channels.

Therefore, to compute $\mathcal{Q}$, we first detect edges in the image and compute the probability that they correspond to a shadow. Then we perform a series of smoothing operations to propagate the appropriate confidence values to pixels in $\mathcal{Q}$ that do not lie on image edges. The form of the final confidence map can be seen in Figure 5.2.

We first apply the Sobel edge detector to the original image I, obtaining a set of gradients, ∇I. We also calculate a set of illumination invariant representations of the original image I. We refer to the k-th illumination invariant representation of I as $I^{(k)}$. An illumination-invariant representation of the original image I will, ideally, not contain any effects of illumination, such as cast shadows and shading [40, 179, 27, 33]. Having such a representation, we can compare the gradients in the original image with gradients in the illumination-invariant representation to attribute the gradient to either shadows/shading or texture. The illumination invariants we chose to use for our experiments are the normalized RGB and $c_1c_2c_3$ representations [41]. We apply the Sobel edge detector to each illumination-invariant image representation $I^{(k)}$ to obtain the corresponding gradients $\nabla I^{(k)}$.

To compute a confidence that each pixel i belongs to a shadow border, we compare the gradients from the original image and each illumination invariant. We define the confidence value for pixel i as:

$$Q(i) = \max_k \left\{ \max \left\{ \|\nabla I(i)\| - w_I^k \left\| \nabla I^{(k)}(i) \right\|, 0 \right\} \right\}. \tag{5.5}$$

Because in practice some gradients related to illumination appear in the illu-

mination invariant representations, we take the maximum of the differences between gradients in the original image and illumination invariants. The weights w_I^k were learned from a training set of images, which was a subset of the dataset of images with hand-annotated cast shadows provided by [200].

After obtaining this initial set of confidences, we apply a smoothing operation for a fixed number of iterations to propagate the confidences to pixels that do not belong to detected image edges. In this smoothing operation, the new confidence value $\hat{Q}(i)$ of pixel i with previous confidence $Q(i)$ is set to be:

$$\hat{Q}(i) = \begin{cases} (1-\lambda)Q(i) + \lambda\overline{Q}(i), & \text{if } \|\nabla I(i)\| < \theta_e \\ \max\{Q(i), \overline{Q}(i)\}, & \text{otherwise} \end{cases}, \tag{5.6}$$

where $\overline{Q}(i)$ is the average of confidence values in a 3x3 neighborhood centered at pixel i. The value of λ was set to 0.5 and θ_e is a small threshold, so that edges with gradient magnitudes less than θ_e are not significant.

Similarly, we create a smoothed version of the edge gradients by setting the new gradient direction of each pixel to be the average of itself and its neighbors, weighted by their relative confidence values. The resulting confidence map and gradient directions can be seen in Fig.5.2.

5.4 Energy minimization

To find the optimal light parameters, we need to minimize the energy $E(\theta_{\mathcal{L}})$ in Eq.5.3. This energy contains multiple local minima, while we also cannot get a good approximation to its gradient. The evaluation of the energy for different parameters, however, is relatively fast. We therefore use a move-making approach, where we start from a random initial set of parameters, and perform a number of iterations, examining at each iteration a random step from the current parameter values:

For the first K iterations, the generation of proposed parameters $\hat{\theta}_{\mathcal{L}}$ is done randomly, to randomly sample the whole parameter space. After the first K iterations, $\hat{\theta}_{\mathcal{L}}$ is generated by choosing the proposed light direction by sampling a von Mises-Fisher distribution centered at the previous estimate of light direction (if we want to estimate intensities at the same time, we also choose a proposed intensity as a sample from a normal distribution around the previous intensity estimate).

If the light intensity is not estimated as part of the energy minimization,

Algorithm 1 Minimization of $E(\theta_{\mathcal{L}})$

```
Light parameters: θ_L ← random parameters
Energy minimum: E_min = E(θ_L)
loop
  generate proposed parameters θ̂_L given θ_L
  if E(θ̂_L) < E(θ_L) then
    E_min ← E(θ̂_L)
    θ_L ← θ̂_L
```

we estimate it afterwards, using the estimate $\hat{\mathbf{d}}_0$ of light direction we obtained. The intensity estimate $\hat{\alpha}_0$ is the median of local intensity estimates along the expected shadow edges:

$$\hat{\alpha}_0 = \text{median}_{i \in \hat{\mathcal{E}}(\theta_{\mathcal{L}}|\mathcal{G})} \left\{ \frac{\overline{I}_{out} - \overline{I}_{int}}{\max\{\mathbf{n}_i \cdot \hat{\mathbf{d}}_0, 0\}} \right\}. \tag{5.7}$$

This very simple approach to minimize the energy $E(\theta_{\mathcal{L}})$ proved effective because it samples the whole parameter space, avoiding many of the local minima, and then concentrates its effort to the area around the best solution so far. However, it would be very desirable in the future to examine other approaches that can give some guarantees about optimality, while also reducing the number of times the energy $E(\theta_{\mathcal{L}})$ has to be evaluated during minimization.

5.4.1 Dealing with multiple light sources

In our discussion so far we have examined only the case of a single light source. When multiple light sources are present, there will be multiple shadow outlines that can be explained by the provided geometry $\mathcal{G}$. We can deal with this case by discovering light sources one-by-one: We estimate the direction and intensity of each light source j, and then remove the corresponding edges from the edge confidence map $\mathcal{Q}$ (removing the corresponding edge pixels and then re-applying the smoothing operation). We then repeat, estimating the next light source from the new, reduced edge confidence map. The process stops when the last discovered light source has very low average confidence values along its projected shadow border, or has near-zero intensity. This procedure can allow not only the estimation of the parameters of multiple light sources, but also to determine the number of light sources illuminating the scene.

5.5 Results

method	NNLS [166]	our method all samples	our method 20% of edges
exact geometry	3.84	1.82	2.06
exact geom.+noise	22.05	1.67	2.03
approximate geom.	13.95	4.00	5.86
approximate geom.+noise	33.69	4.76	6.28

Table 5.1: Average error in light direction estimation for a set of synthetic images. The images were rendered using 1 known point light source, and the displayed error is the angle between the real and estimated light directions, in degrees. We compare with the non-negative least squares optimization (NNLS) approach proposed in [166]. Examples of the images and geometry used are shown in Fig.5.3. The second row shows results with our approach when all expected shadow border pixels are used to evaluate the energy, and the third row shows results with our approach when only 20% of expected shadow border pixels is used, achieving a 5-fold speedup with small deterioration of the results. Our approach significantly outperforms [166] and it is influenced much less than [166] by noisy shadow input and coarse knowledge of geometry.

We evaluated our approach quantitatively using a synthetic dataset of 3D models rendered with a known distant point light source, as well as qualitatively with images collected from Flickr [129]. A total of 1000 iterations was performed for each image.

In the results we present here, we compare with the approach of [166], which is a well-established approach for illumination estimation based on non-negative least squares optimization (explained in Chapter 2), and with the MRF approach of [129], which is an earlier version of the work of ours presented in the next chapter. The MRF approach of the next chapter [129] is a hybrid approach, in the sense that it combines both information from pixel intensities and from shadow edges. It includes a term that incorporates the idea we presented in this chapter, which in the MRF formulation of the next chapter is referred to as the *shadow shape-matching prior*. Please notice that only in

figure 5.6 we compare the approach of this chapter to the MRF approach of the next chapter [129] *without* the shadow shape-matching prior, to clearly demonstrate the benefit of taking advantage of shadow edge information.

Results on the synthetic dataset are shown in Table 5.1. Examples of the synthetic images and models used are shown in Fig.5.3. The direction and intensity of the light source was chosen randomly. We examined four different cases:

Exact geometry: We used the same 3D model to render the image and to estimate illumination.

Approximate geometry: We used a 3D model that coarsely approximated the original geometry by a bounding box and a ground plane to estimate illumination.

Exact geometry and noisy shadow input: We used the same 3D model to render the image and to estimate illumination, as above, and a noisy initial shadow estimate. The latter was obtained by adding random dark patches to the rendered shadow (Fig.5.3). We used this form of noise because, on one hand our methods are relatively insensitive to spatially-uniform random noise, and on the other, in real data the errors generally affect large image regions which get mislabeled, which is emulated by this patch-based noise.

Approximate geometry and noisy shadow input: We estimated illumination parameters using a coarse 3D model and a noisy initial shadow estimate, as described above.

Table 5.1 shows the average error in light direction estimation, in degrees. It is easy to notice that our approach is almost unaffected by errors in the initial shadow estimate, which have been simulated by the noisy shadow input. The non-negative least squares optimization approach of [166] on the other hand exhibits a significant drop in accuracy when noise is present in the shadow estimate, as well as in the case the knowledge of scene geometry is approximate. Figure 5.4 shows the convergence of our approach in the case of synthetic data.

Figure 5.5 shows results with our approach on a set of images of cars from Flickr. The geometry in these images consists of a ground plane and a 3D bounding box representing the car. We compare our results with the results obtained in [129], which is an earlier version of the work we present in Chapter 6. The estimated illumination is shown by rendering a synthetic orange sun dial, illuminated by the illumination estimate obtained with our method, into the original image. Although accurate comparisons are difficult since there is no ground truth for illumination in these images, a visual inspection shows that our results are equally convincing than those obtained by the MRF ap-

Figure 5.3: Examples of synthetic images used for quantitative evaluation. Image size was 100x100 pixels. From left to right: a) the full-resolution 3D model, rendered with 1 point light source; b) the full-resolution 3D model, after the addition of noise, as described in the text (noise is random and may coincide with the shadow); c) the approximate 3D model corresponding to the full-resolution model on the left, rendered under the same illumination (for demonstration; this model is used for the illumination estimation only, in the experiments with approximate geometry).

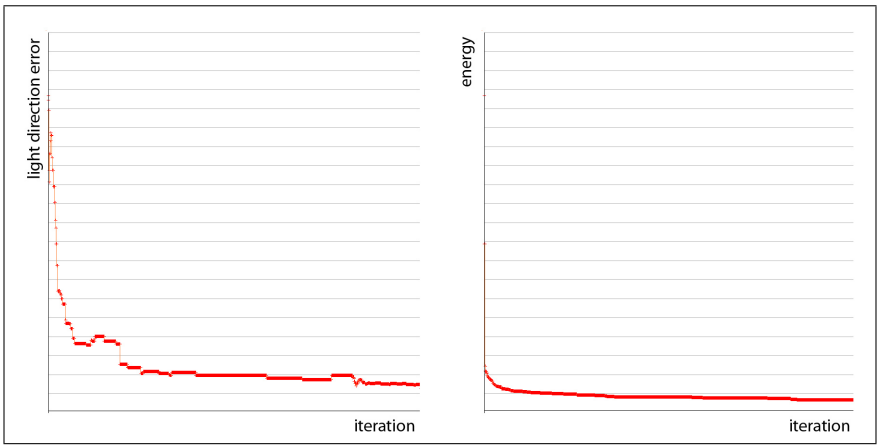

Figure 5.4: Convergence of our algorithm. Left: the error in the estimated light direction, averaged over a set of synthetic examples, per iteration; right: the average energy per iteration.

proach of [129]. On the other hand, the approach of [129] is significantly more computationally demanding (Table 5.2).

Figure 5.7 demonstrates the flexibility of our method with regard to shadow detection. In this case, we compare the illumination estimate obtained when using all image edges (obtained with a Sobel detector) compared to using only potential shadow edges (by utilizing illumination invariants as described earlier). Our approach can select those image edges that correspond to plausible

Figure 5.5: Results with images of cars from Flickr. The results of illumination estimation are presented by rendering a synthetic orange sundial to the original image, using the estimated illumination. Top: the results with our method; bottom: results with the much more computationally intensive method proposed in [129]. Our results are equivalent or better than the results from [129], although our method uses only simple shadow edge detection and a much more efficient optimization to estimate illumination parameters.

cast shadows, and obtain a good illumination estimate, even when no initial shadow edge detection is performed.

Examples of the 3D geometry used for illumination estimation in the case of Flickr images is shown in figure 5.8.

Excluding the cost of raytracing shadows, the computational cost of our method is linear to the number of edge samples; the number of edge samples used can be reduced without significant impact to the final results, in order to improve performance. In our unoptimized implementation, on a system with an Intel i5 CPU, each iteration took 4-10msec depending on image size, or 1-3msec when using one of every 5 shadow edge samples (in the case of synthetic results in Table 5.2). Paired with a GPU raytracer, the total time for our algorithm can be limited to 5-60 seconds per image, which is a significant advantage compared to computationally intensive methods such as the more complete illumination estimation method we present in Chapter 6. As our results show, this improvement usually comes at little cost to the accuracy of the estimated results.

One significant advantage of the approach proposed here is that potential shadow edges which are far from the shadow edges generated by the known geometry are not penalized at all. Therefore, our method will ignore real shadows that cannot be explained by the geometry, if the geometry models only

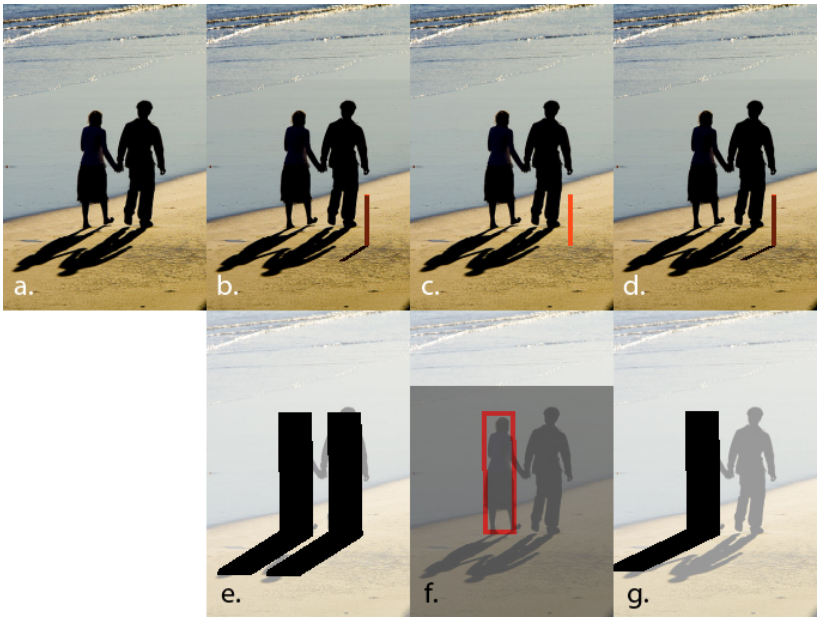

Figure 5.6: Advantages of our approach: in this figure we compare the behavior of our method with [129] when the geometry is only partially modeled. a) original image from Flickr [128], depicting 2 persons; b) the illumination estimation result when both people are modeled, with our method (illustrated by rendering an orange sundial into the original image with the estimated illumination).; c) illumination estimate with [129] (without the shadow shape-matching prior) when *only one* of the two people is modeled. The algorithm tries to explain both shadows with one object, resulting in a light source placed under the scene; d) our approach using **the same** 3D model as in (c), when *only one* of the two people is modeled - the illumination estimate is convincing and almost the same as in (b) where full geometry was given. In (e), (f) and (g) we show the 3D model used to estimate illumination in (b),(c) and (d) respectively, rendered with the estimated illumination. Notice that (f) and (g) show **the same** 3D model, but because the estimated light is under the scene in (f), we marked the model with a red outline to make it visible.

small part of the actual scene. At the same time, wrongly detected shadow edges need not be accounted for, and have no effect on the energy of the final solution. This advantage of our approach is demonstrated in figure 5.6. In this example only one of two nearby objects is modeled. For comparison, we present the result using the MRF approach of [129] (see next chapter) without the shadow shape prior term,as mentioned earlier. The partial modeling

method	running time (sec)
[129] (MRF)	244
our method, all samples	4.4
our method, 20% of samples	1.2

Table 5.2: Running times for our algorithm compared to the state-of-the-art approach of [129], for a 500x300 pixel image. The times *exclude* the time spent ray-tracing (which is the same for both approaches and can be reduced to less than 1 sec using hardware acceleration).

of the scene causes the modified MRF approach of [129] (and probably most approaches that are based on an error computed over all shadow pixels) to try to explain all shadows using the provided geometry, resulting in erroneous illumination estimation. Our approach, on the other hand, correctly estimates illumination by associating it only with a subset of the observed shadow edges. One drawback is that this kind of approach could potentially ignore real shadows when the geometry differs substantially, but in our experiments modeling objects with 3D bounding boxes, even with inaccurate modeling, was enough to associate the illumination solution with the correct set of shadow edges.

We only used light sources that produce sharp shadows in our experiments. A limitation of the proposed approach is that it cannot handle well soft shadows produced by area light sources.

5.6 Conclusions

In this chapter, we presented an approach to estimate illumination from a subset of shadow borders. The advantages of this approach, as we demonstrated, are that illumination estimation relies much less in the quality of shadow detection, while at the same time allowing the partial and coarse modeling of the 3D geometry of the scene. Our results show that our approach can estimate illumination even when no shadow detection is performed (Fig.5.7), since the sets of image edges that match potential shadow outlines are limited. The accuracy of our results is better than previous approaches [166], and comparable to the method of [129], while achieving relatively low computational complexity, which can be further controlled by limiting the number of edge samples used in our computations as desired. In the future, we would like to examine optimization approaches that can give us guarantees on the optimality of the

Figure 5.7: Results comparing illumination estimation based on potential shadow edges (left) and all edges in the image (right). It is clear that our method can work even when we do not explicitly detect shadow edges, because few image edges match potential shadow silhouettes. The results of illumination estimation are presented in the bottom by rendering a synthetic orange sundial to the original image, using the estimated illumination. Maps $\mathcal{Q}$ are in the top row. This is an image where we failed to obtain any meaningful illumination estimate with the method of [129].

solution and to examine other ways to build the shadow edge confidence map and the corresponding map of edge orientations, for example through a diffusion process. Another exciting direction for future work would be to associate the matching of the synthetic shadow silhouette with deformable geometric models, allowing some refinement of geometry concurrently with illumination estimation.

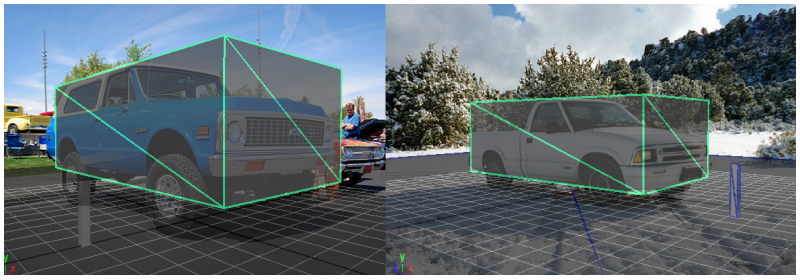

Figure 5.8: The geometry used to approximate the cars in images from Flickr. The geometry consists of a bounding box (green) that encloses the body of the car, and a plane for the ground. Camera parameters were selected by hand to match each scene.

Chapter 6

Scene Photometry in a Global MRF Model

In the previous chapters we discussed two different approaches towards the estimation of illumination under approximate knowledge of geometry and initial shadows. As discussed in the introduction, one of our goals is to be able to utilize illumination when provided with images of complex, real scenes, in order to extract more information about the scene, or integrate illumination in more complicated tasks. Towards this goal, in this chapter we present a new model to capture the interaction of illumination, geometry and cast shadows in an image. This model combines ideas from the approaches described in the two previous chapters in a complete framework. It effectively captures every aspect of the problem and, through a hybrid optimization scheme, is able to infer *all three components* of the problem (shadows, illumination, geometry), given some initial hints. Through this framework, not only are very coarse approximations of geometry, such as bounding boxes, enough to estimate illumination, but geometry reasoning can be incorporated with illumination estimation. In the following sections we show how, for example, the geometry of the occluders can be refined as part of the illumination estimation process. The initial approximate geometric information we require could be derived as part of more general scene understanding techniques, while enabling illumination estimation to be incorporated in the scene understanding loop; the obtained illumination and geometry information could be a crucial contextual prior in addressing various other scene understanding questions.

Graphical models can efficiently incorporate different cues within a unified framework [186]. Hence, in order to deal with the complex illumina-

tion/geometry/shadows estimation problem robustly in a flexible and extensible framework, we jointly model the geometry, light sources, and shadow values within an Markov Random Field model, and all the latent variables can then be simultaneously inferred through the minimization of the energy of the MRF. To the best of our knowledge, this is the first time that the interaction of illumination and geometry in the image formation is addressed using an MRF model.

The MRF model we propose captures the interaction between geometry and light sources and combines it with image evidence of cast shadows. Cast shadow detection is well-posed in terms of graph topology, since it can be expressed using a graph in the form of a 2-dimensional 4-connected lattice, where each image pixel corresponds to a graph node. Modeling in the MRF model the creation of cast shadows from the interaction of light sources and geometry, on the other hand, implies a potential dependence between each pixel and all nodes representing the light sources and the occluder geometry. This generally results in higher-order cliques in the graph representing our MRF model. Further complications arise by the fact that the number of light sources is unknown, resulting in unknown MRF topology, and the search space is continuous. In our model, we are able to reduce the search space and identify the MRF topology through an initial illumination estimate obtained using a voting algorithm. To tackle inference in the resulting higher-order MRF model with both discrete and continuous variables, we describe two methods based on discrete optimization.

We make the following *assumptions* (common in illumination modeling):

- an initial coarse 3D geometry is known,
- the illumination environment can be approximated by a set of distant light sources,
- the reflectance of surfaces is roughly lambertian.

Futhermore, when we discuss how occluder geometry parameters can be estimated jointly with illumination, we assume that these occluders are identified in the original image by providing a 2D bounding box, and one or more candidate geometric models that could potentially approximate the object 3D geometry (see Fig.6.1).

In the end of this chapter we provide qualitative evaluation of the proposed method on different datasets, including images captured in a controlled environment, car images collected from Flickr and images from the Motorbikes

class of Caltech 101 [110]. We also provide quantitative results on a synthetic dataset. The experimental evaluation shows that our method is robust enough to be able to use geometry consisting of bounding boxes or a common rough 3D model for a whole class of objects, while it can also be applied to scenes where some of our assumptions are violated. Results on geometry parameter estimation show that through our model we can extract useful information about object geometry and pose from the cast shadows.

The remainder of this chapter is organized as follows: Sec. 6.1 presents related fundamentals; Sec. 6.2 describes the MRF model to jointly estimate the shadows, illumination and geometry parameters. In Sec. 6.3 we discuss the inference process. Experimental evaluation is provided in Sec. 6.4, and Sec. 6.5 concludes this chapter.

6.1 Fundamentals

We follow the same model as described in the beginning of Section 2.1.1 in Chapter 2. We adopt a commonly used set of assumptions: the surfaces in the scene exhibit lambertian reflectance, and the scene is illuminated by point light sources at infinity, as well as some constant ambient illumination term. Under these assumptions, the outgoing radiance at a pixel i can be expressed as:

$$L_o(\mathbf{p}) = \rho_{\mathbf{p}} \left(\alpha_0 + \sum_{i=1}^{N} V_{\mathbf{p}}(\mathbf{d}_i) \alpha_i \max\{\mathbf{d}_i \cdot \mathbf{n}_{\mathbf{p}}, 0\} \right), \tag{6.1}$$

where N is the number of light sources, $\rho_{\mathbf{p}}$ is the albedo at point $\mathbf{p}$, α_0 is the ambient intensity, $\alpha_i, i \in \{1, ..., N\}$ is the intensity of the i-th light source, $\mathbf{d}_i$ is the illumination direction of the i-th light source, and $V_{\mathbf{p}}(\mathbf{d}_i)$ is a visibility term for direction $\mathbf{d}_i$ at point $\mathbf{p}$.

Therefore illumination information is fully captured by parameters $\theta_{\mathcal{L}} = \{\alpha_0, \alpha_1, ..., \alpha_N, \mathbf{d}_1, ..., \mathbf{d}_N\}$.

We first obtain an initial cast shadow estimate from the input image $\mathbf{I}$. This estimate should contain the shading intensity at each pixel in shadow, without any variations due to albedo ρ, and the non-shadow pixels of $\mathbf{I}$ should be masked out. Ideally, therefore, the value of each shadow pixel (x, y) in such a shadow image $\mathbf{I}_s$ would be the shading at that point due to the non-occluded

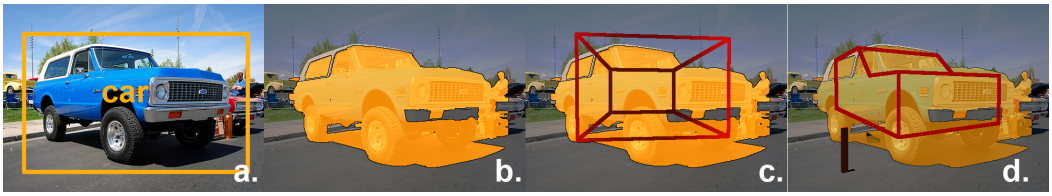

Figure 6.1: Intermediate steps for geometry parameter estimation: a) Input: the original image, with a 2D bounding box localizing the object and a label indicating which are the candidate geometric models to explain this object. b) The mask obtained from the 2D bounding box using GrabCut. c) A 3D bounding box is positioned randomly inside the 2D bounding box. d) After the execution of our method, we obtain the most probable geometry to explain the object, an estimate of 3D pose parameters, as well as an estimate of illumination and shadows.

light sources, given by:

$$I_s(x,y) = \alpha_0 + \sum_{i=1}^{N} V_{\mathbf{p}}(\mathbf{d}_i)\alpha_i \max\{\mathbf{d}_i \cdot \mathbf{n}_{\mathbf{p}}, 0\}, \tag{6.2}$$

where p is the 3D point where (x, y) projects to. In practice we can obtain a cast shadow estimate $\hat{\mathbf{I}}_s$ which is a rough approximation of $\mathbf{I}_s$.

In our experiments, we use the approach we presented in Chapter 3 in this thesis in order to obtain the shadow estimate $\hat{\mathbf{I}}_s$. The initial shadow estimate $\hat{\mathbf{I}}_s$ we obtain through this method is used as input to the MRF model we present later in this chapter. The final shadow estimate produced by our MRF inference process attempts to remain close to this initial estimate, as well as conform to the synthetic shadow expected by the estimated illumination and geometry configuration.

6.1.1 Geometry modeling

One of the goals of this work is to provide a model that allows reasoning about illumination to be incorporated in more complex scene understanding tasks. Towards this goal, we describe here how we can incorporate objects with unknown parameters to be estimated to our model. Estimation of these parameters happens jointly with the estimation of illumination and cast shadows. Different parametrizations of the scene geometry could be handled by our

model without significant changes, as long as the total number of geometry parameters remains small.

As mentioned, $\mathcal{G}$ is the known, approximate 3D geometry which is provided as input. We assume that there may also exist a (small) set of objects $\mathcal{O}$, which are the parametric objects we want to estimate. The information we assume as known about the objects $\mathcal{O}$ is restricted, for each object i, to a 2D bounding box that bounds the object in the image, and a set ${\mathcal{G}_{\mathcal{O}}}^{(i)}$ of potential approximate 3D models for this object. The potential 3D models can be thought as the geometric models representing common instances of the class to which object i belongs (e.g. if the object is a car, we could assume a small number of 3D models representing common car shapes). Our goal is to recover, concurrently with illumination estimation, the most probable geometry for each of these objects, as well as the most probable orientation, translation and scale for each of them, in order to best approximate the real scene geometry.

In figure 6.1 we show a visual representation of some of the input required by our method in order to perform geometry estimation, as well as of the intermediate steps to obtain the final estimate of parameters.

In the following sections we present a model to jointly estimate the shadows, the illumination parameters $\theta_{\mathcal{L}}$ and a set of geometry parameters from the approximate shadow cue $\hat{\mathbf{I}}_s$.

6.2 Global MRF for Scene Photometry

We associate the image-level evidence for cast shadows with high-level information about geometry and the light sources through the MRF model described in this section.

6.2.1 Markov Random Field Formulation

The proposed MRF consists of one node for each image pixel $i \in \mathcal{P}$, one node for each light source $l \in \mathcal{L}$, one node for the ambient intensity α_0 and one node for the geometry of each object k in the set of objects $\mathcal{O}$. Each pixel node, all the light nodes and all the object nodes compose a high-order clique $c \in \mathcal{C}$. The 4-neighborhood system [17] composes the edge set $\mathcal{E}$ between pixels. The

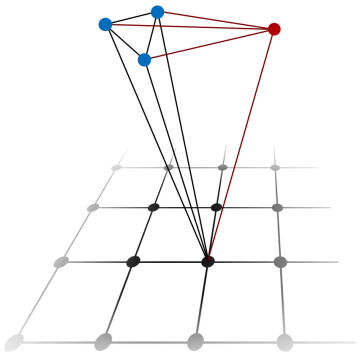

Figure 6.2: MRF topology: This is part of the graph representing an example MRF model with 2 light sources and one geometry node. The light source nodes are shown in blue on the top, with 2 nodes representing light sources and a special node representing the ambient illumination. The red node corresponds to a parametrized geometric model. The nodes in black are the pixel nodes, connected in a 2D lattice. Each pixel node is connected with all light and geometry nodes (shown here only for one pixel node for simplicity).

energy of our MRF model has the following form:

$$\begin{aligned} E(\mathbf{x}) &= \sum_{i\in\mathcal{P}} \phi_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \psi_{i,j}(x_i, x_j) + \sum_{k\in\mathcal{O}} \phi_k(x_k) \\ &\quad + \sum_{l\in\mathcal{L}} \phi_l(x_l, \mathbf{x}_\mathcal{O}) + \sum_{i\in\mathcal{P}} \psi_c(x_i, \mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O}), \end{aligned} \tag{6.3}$$

where $\phi_i(x_i)$ and $\phi_k(x_k)$ are the singleton potentials for pixel nodes and object nodes respectively, $\psi_{i,j}(x_i, x_j)$ is the pairwise potential defined on a pair of neighboring pixels, $\phi_l(x_l, \mathbf{x}_\mathcal{O})$ is the clique potential expressing a shadow shape-matching prior, and $\psi_c(x_i, \mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O})$ is the high-order potential associating all lights in $\mathcal{L}$, all objects in $\mathcal{O}$ and a pixel x_i.

The latent variable x_i for pixel node $i \in \mathcal{P}$ represents the intensity value for that pixel. We uniformly discretize the real intensity value $[0, 1]$ into N bins to get the candidate set $\mathcal{X}_i$ for x_i. The latent variable x_l for light node $l \in \mathcal{L}$ is composed of the intensity and the direction of the light. We sample the space in the vicinity of the light configuration obtained by the previous voting approach to get the candidate set $\mathcal{X}_l$ for x_l (see details later in this section).

By $\mathbf{x}_{\mathcal{O}}$ we signify the labels corresponding to the objects in $\mathcal{O}$. The label x_k^O of object node k determines a set of parameters $(g_k, \phi_k, t_x, t_y, t_z, s_x, s_y, s_z)$, where g_k is an index into ${\mathcal{G}_{\mathcal{O}}}^{(k)}$ that determines which of the potential object geometries is selected for label x_k^O, ϕ_k is the azimuth orientation of the object, (t_x, t_y, t_z) is the translation and (s_x, s_y, s_z) is the scale of the object.

Figure 6.2 shows the topology of the MRF. In particular, it shows the MRF nodes for an example with 2 light sources and one geometry node, and also the edges connecting the pixel nodes and one of the pixel nodes with the light and geometry nodes.

Singleton Potentials for Pixel Nodes

This term encodes the similarity between the estimated intensity value at pixel i and the shadow cue value $\hat{I}_s(i)$ and is defined as:

$$\phi_i(x_i) = w_s \min\{\left| x_i - \hat{I}_s(i) \right|, t_p\}. \tag{6.4}$$

where an upper bound t_p for this cost term is used to avoid over-penalizing outliers and w_s is a positive weight coefficient (same for w_l, w_p and w_c below).

Singleton Potentials for Geometry

In our attempt to extract information about the geometry of object k, in the model of Eq.6.3 we obviously take into account the information in the shadow cast by object k. However, the cast shadow provides only one projection of the object, which is often insufficient to extract useful information about the object shape. We can, however, obtain a second projection of the object, the one onto the image plane, which will provide us with extra information to make reasoning about the object pose and shape possible.

To obtain the shape of the object on the image plane, we use GrabCut [157] with the user-provided 2D bounding box for the object as input. GrabCut gives us a foreground/background segmentation, where pixels in the foreground $\mathcal{F}$ are the pixels most likely to belong to the object contained in the initial 2D bounding box.

The singleton potentials $\phi_k(x_k)$ penalize geometry labels x_k that are inconsistent with the extracted shape $\mathcal{F}$ of the object k in the image. This potential also penalizes geometry labels x_k that correspond to a scale that significantly

deforms the initial geometry. The form of the potential is:

$$\phi_k(x_k) = \sum_{i \in \mathcal{P}} \left(\mathcal{F}(i) - \mathcal{M}_{x_k}(i)\right)^2 + w_s \left\| \mathbf{x}_k^{(scale)} - [1,1,1] \right\|_2, \tag{6.5}$$

where $\mathbf{x}_k^{(scale)}$ is a vector (s_x, s_y, s_z) determining the object scale corresponding to label x_k, $\mathcal{F}$ is the object mask obtained by GrabCut:

$$\mathcal{F}(i) = \begin{cases} -1 & \text{if } i \in \text{background} \\ +1 & \text{if } i \in \text{foreground} \end{cases} \tag{6.6}$$

and $\mathcal{M}$ is the mask corresponding to the projection $I_k^{\mathcal{O}}$ of the geometry assigned to object k from label x_k, at the corresponding rotation, translation and scale:

$$\mathcal{M}(i) = \begin{cases} -1 & \text{if } i \in I_k^{\mathcal{O}} \\ +1 & \text{if } i \in I_k^{\mathcal{O}} \end{cases} . \tag{6.7}$$

As demonstrated in our experiments (Fig.6.14), the obtained mask $\mathcal{M}$ is not by itself adequate for determining the geometry parameters. The combination of the mask $\mathcal{M}$ with the information contained in shadow regions in our MRF model, however, allows us to obtain a good estimate of the geometry parameters.

Pairwise Potentials

We adopt the well-known *Ising* prior [39] to define the pairwise potential between neighboring pixels $(i,j) \in \mathcal{E}$ to favor neighboring pixels having the same value:

$$\psi_{i,j}(x_i, x_j) = \begin{cases} w_p & \text{if } x_i \neq x_j \\ 0 & \text{if } x_i = x_j \end{cases} \tag{6.8}$$

Shadow Shape-matching Prior

Terms $\phi_l(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})$ incorporate into the MRF model a *shadow shape-matching prior* for light l, in order to favor illumination and geometry configurations generating shadow shapes that match observed shadow outlines. The shadow shape-matching prior implements the idea presented in more detail in Chapter 5. It evaluates the quality of the matching between the observed edges in the image and the edges expected given a light configuration $\mathbf{x}_l$ for light l and geometry configuration $\mathbf{x}_{\mathcal{O}}$.

We take a slightly different approach than that of the previous chapter in the way we compute the edge maps used to penalize the shadow shapes. The way presented here is coarser, and is based on producing different distance maps for a small number of discretized edge directions. Then the direction of each edge generated by the light and geometry configuration is penalized based on the distance maps corresponding to the two closest discretized directions. This results in a slightly weaker (more approximate) constraint than that of the work in the previous chapter. It, however, appeared to give better results when used as a component of our MRF model. Notice that due to the modularity of the MRF formulation, we could simply substitute this form of the shape-matching prior with the approach of the previous chapter without any other modifications in the MRF modeling or inference.

During the initialization phase of our algorithm, we first apply gaussian smoothing and the Sobel edge detector [45] to detect edges in the shadow cue image. Let $\tau(i) \in [0, 2\pi)$ be the angle of the gradient at pixel i with the x-axis, and $\hat{\tau}(i) \in \{0, K-1\}$ a quantization of $\tau(i)$. For each possible direction $d \in \{0, K-1\}$, we compute a distance map υ_d so that, for pixel i, $\upsilon_d(i)$ is the distance from pixel i to the closest edge pixel of orientation d.

During inference, for pixel i with gradient angle $\tau(i)$, the distance function is computed by interpolating between the distance map values for the two closest quantized orientations:

$$dist_{\tau(i)}(i) = (1-\lambda) \cdot \upsilon_{\hat{\tau}(i)}(i) + \lambda \cdot \upsilon_{\hat{\tau}(i)+1}(i), \tag{6.9}$$

$$\lambda = \left\{ \frac{K \cdot \tau(i)}{2\pi} \right\}, \tag{6.10}$$

where $\{.\}$ indicates the fractional part. In our experiments, we chose $K = 4$.

The shape-matching prior expresses the quality of the match between the observed edges in the shadow cue image and the edges of the synthetic shadow $\mathcal{S}_l$ associated with $\mathbf{x}_l$ and geometry configuration $\mathbf{x}_\mathcal{O}$:

$$\phi_l(\mathbf{x}_l, \mathbf{x}_\mathcal{O}) = w_l \frac{1}{|\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_\mathcal{O})|} \sum_{i \in \mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_\mathcal{O})} dist_{\tau_{\mathcal{S}_l}(i)}(i), \tag{6.11}$$

where $\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_\mathcal{O})$ is the set of all pixels that lie on edges of the shadow $\mathcal{S}_l$ generated by light label $\mathbf{x}_l$ and $\tau_{\mathcal{S}_l}(i)$ is the gradient angle of the synthetic shadow edge generated by x_l at pixel i. To determine the set of shadow edge pixels $\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_\mathcal{O})$, we generate the shadow $\mathcal{S}_l$ created by light label $\mathbf{x}_l$ and the

geometry $\mathbf{x}_{\mathcal{O}}$ and then apply gaussian smoothing and the Sobel edge detector. The set $\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})$ contains all pixels whose gradient magnitude is above θ_e.

Higher-order Potentials

The higher-order terms $\psi_c(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}})$ impose consistency between the light source labels $\mathbf{x}_{\mathcal{L}}$, the geometry labels $\mathbf{x}_{\mathcal{O}}$ and the pixel intensity values.

Let $\mathcal{S}$ be the synthetic shadow, generated by light configuration $\mathbf{x}_{\mathcal{L}}$ and geometry configuration $\mathbf{x}_{\mathcal{O}}$. The intensity at pixel $i \in \mathcal{S}$ is:

$$s'_i(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) = \mathbf{x}^{\alpha_0} + \sum_{l \in \mathcal{L}} x_l^{\alpha} V_i(\mathbf{x}_l^{dir} | \mathbf{x}_{\mathcal{O}}) \max\{-\mathbf{x}_l^{dir} \cdot \mathbf{n}(i), 0\},$$

where $\mathbf{x}^{\alpha_0}$ corresponds to the ambient intensity, x_l^{α} is the light intensity component of x_l, $\mathbf{x}_l^{dir}$ is the light direction component, $\mathbf{n}(i)$ is the normal at 3D point $\mathbf{p}$ imaged at pixel i and $V_i(\mathbf{x}_l^{dir}) \in \{0, 1\}$ is the visibility term for light direction $\mathbf{x}_l^{dir}$ at 3D point $\mathbf{p}$ (cf. Eq.2.3). For pixels $i \notin \mathcal{S}$, we set $s'_i(\mathbf{x}_{\mathcal{L}}) = 1$, according to the definition of our shadow cue $I_s(i)$. The clique potential is defined as:

$$\psi_c^{(1)}(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) = w_c \min\{(s'_i(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) - x_i)^2, t_c\}, \tag{6.12}$$

where t_c is also an upper bound to avoid over-penalizing outliers.

In cases where the geometry $\mathcal{G}$ is far from the real scene geometry, a light configuration that does not generate any visible shadows in the image might result to a lower MRF energy than the true light source. To avoid this degenerate case, we introduce the term $\psi_c^{(2)}(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}})$, which assignes a fixed penalty to light configurations that do not generate any visible shadows in the image. The final form of the clique potential is:

$$\psi_c(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) = \psi_c^{(1)}(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) + \psi_c^{(2)}(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}). \tag{6.13}$$

6.2.2 Initializing the MRF Model

As mentioned earlier, the continuous search space complicates inference in our MRF model. Furthermore, in our discussion of the model so far, we assumed that the number of light sources $|\mathcal{L}|$ is known. In practice, however, $|\mathcal{L}|$ may be unknown, which results in unknown MRF topology. To deal with these two issues, we use a rough initial illumination estimate both to determine $|\mathcal{L}|$, if it

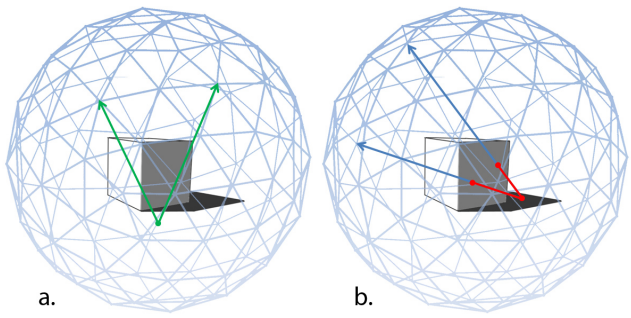

Figure 6.3: Our voting algorithm: a) pixels that are not in shadow vote for all possible illumination directions that are not occluded; b) pixels in shadow vote for all illumination directions that are occluded by the geometry.

is unknown, and to set the initial values of the light source variables, before inference begins.

To obtain this rough illumination estimate we use the greedy approach described in Algorithm 2, based on the shadow cue $\mathbf{I}_s$ and geometry $\mathcal{G}$. We examine a fixed set of possible illumination directions, corresponding to the nodes of a geodesic sphere [166]. In each iteration of this algorithm, the pixels in shadow, which are not explained by already discovered light sources, vote for all occluded illumination directions. Pixels not in shadow vote for all illumination directions that are not occluded (see Fig.6.3 for an example). After all pixels cast a vote, the most popular direction is chosen as the direction of the new light source. Having the light source direction, we estimate the light source intensity using the median of local intensity estimates from each pixel in the shadow of this light source, and the new light source is added to the set of discovered light sources. The algorithm stops when the estimated intensity of the new light source is near zero, meaning that it doesn't have a significant contribution to the observed shadows.

6.3 Inference

We simultaneously estimate the cast shadows, illumination and geometry parameters by minimizing the MRF's energy defined in Eq. 6.3:

$$\mathbf{x}^{opt} = \arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{6.14}$$

Algorithm 2 Voting for initial illumination estimate

Lights Set: $\mathcal{L} \leftarrow \varnothing$
Direction Set: $\mathcal{D} \leftarrow$ all the nodes of a unit geodesic sphere
Pixel Set: $\mathcal{P} \leftarrow$ all the pixels in the observed image
loop
  votes[$\mathbf{d}$] $\leftarrow 0, \forall \mathbf{d} \in \mathcal{D}$
  for all pixel $i \in \mathcal{P}$ **do**
    for all direction $\mathbf{d} \in \mathcal{D} \setminus \mathcal{L}$ **do**
      if $I_s(i) < \theta_S$ **and** $\forall \mathbf{d}' \in \mathcal{L}, V_i(\mathbf{d}') = 0$ **then**
        if $V_i(\mathbf{d}) = 1$ **then** votes[$\mathbf{d}$] $\leftarrow$ votes[$\mathbf{d}$] $+ 1$
      else
        if $V_i(\mathbf{d}) = 0$ **then** votes[$\mathbf{d}$] $\leftarrow$ votes[$\mathbf{d}$] $+ 1$
  $\mathbf{d}^* \leftarrow \arg\max_{\mathbf{d}}(votes[\mathbf{d}])$
  $\mathcal{P}_{\mathbf{d}^*} \leftarrow \{i | c_i(\mathbf{d}^*) = 1 \text{ and } \forall \mathbf{d} \neq \mathbf{d}^*, c_i(\mathbf{d}) = 0\}$
  $\alpha_{\mathbf{d}^*} \leftarrow median \left\{ \frac{1 - I_s(i)}{\max\{-\mathbf{n}(\mathbf{p}(i)) \cdot \mathbf{d}^*, 0\}} \right\}_{i \in \mathcal{P}_{\mathbf{d}^*}}$
  if $\alpha_{\mathbf{d}^*} < \epsilon_\alpha$ **then**
    stop the loop
  $\mathcal{L} \leftarrow \mathcal{L} \cup (\mathbf{d}^*, \alpha_{\mathbf{d}^*})$

Minimizing this energy, however, is challenging, because our MRF model contains high-order cliques of size up to $|\mathcal{L}| + |\mathcal{O}| + 2$.

A straightforward way to minimize the model energy is the high-order clique reduction technique proposed in [65]. This method performs inference in a higher-order MRF with binary labels by reducing any pseudo-Boolean function to an equivalent quadratic one while keeping the minima of the resulting function the same as the original. Like [65], we extend this method to deal with multi-label MRFs by employing the fusion-move [107] and QPBO [52, 86] algorithms. Therefore, a number of iterations is performed, and for each iteration, the algorithm fuses the current labeling L_{cur} and a proposed labeling L_{prop} by minimizing a pseudo-Boolean energy [65].

In our experiments, however, this method failed to provide good solutions (Table 6.1). This can be explained by the complexity of the graph-structure, the large number of labels and the nature of pair-wise and higher order interactions.

To address this failure and efficiently perform inference, we can split the minimization of the energy in Eq.6.3 in two stages [19]. If we assume that the light parameters are fixed, the high-order clique potentials $\psi_c^{(1)}$ in Eq.6.12,

which are part of ψ_c, become singleton potentials of the form:

$$\psi_c^{(1)}(x_i|\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O}) = w_c \min\{(s_i'(\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O}) - x_i)^2, t_c\}. \tag{6.15}$$

This way, for a fixed light configuration $\mathbf{x}_\mathcal{L}$ and a fixed geometry configuration $\mathbf{x}_\mathcal{O}$, after we split ψ_c in $\psi_c^{(1)}$ and $\psi_c^{(2)}$ as in Eq.6.13, we can rewrite the energy of the MRF model in Eq.6.3 as:

$$E(\mathbf{x}) = E_I(\mathbf{x}|\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O}) + E_L(\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O}) + E_G(\mathbf{x}_\mathcal{O}), \tag{6.16}$$

where

$$E_I(\mathbf{x}|\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O}) = \sum_{i \in \mathcal{P}} \left(\phi_i(x_i) + \psi_c^{(1)}(x_i|\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O})\right) + \sum_{(i,j) \in \mathcal{E}} \psi_{i,j}(x_i, x_j) \tag{6.17}$$

is the component of the MRF's energy involving only pairwise potentials, associating the (fixed) light configuration $\mathbf{x}_\mathcal{L}$ and geometry configuration $\mathbf{x}_\mathcal{O}$ with per-pixel variables, and

$$E_L(\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O}) = \sum_{l \in \mathcal{L}} \left(\phi_l(x_l) + \psi_c^{(2)}(\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O})\right), \tag{6.18}$$

$$E_G(\mathbf{x}_\mathcal{O}) = \sum_{k \in \mathcal{O}} \phi_k(x_k) \tag{6.19}$$

are the energy terms associated with the (fixed) light configuration $\mathbf{x}_\mathcal{L}$ and the (fixed) geometry configuration $\mathbf{x}_\mathcal{O}$ but independent of the per-pixel variables.

For a given light configuration $\mathbf{x}_\mathcal{L}$ and geometry configuration $\mathbf{x}_\mathcal{O}$, the energy $E_I(\mathbf{x}|\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O})$ can be minimized using any inference algorithm for pairwise MRF models. The speed of the chosen algorithm is, however, important, because the energy $E_I(\mathbf{x}|\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O})$ is minimized many times (for different light and geometry configurations). To achieve better performance, we used the FastPD algorithm [94] in our experiments.

The energy minimum $\min_x\{E_I(\mathbf{x}|\mathbf{x}_\mathcal{L}, \mathbf{x}_\mathcal{O})\}$ changes with different light configurations and different geometry configurations, as in the simple example shown in Fig.6.4. To minimize $E(\mathbf{x})$, a (blocked) coordinate descent approach in the light and geometry parameter domain is used:

Let $\hat{\mathbf{x}}_\mathcal{L}^{(s-1)}$, $\hat{\mathbf{x}}_\mathcal{O}^{(s-1)}$ be the set of light and geometry parameters that correspond to the minimum energy encountered up to iteration $s-1$. At iteration s, we generate proposed light labels $\mathbf{x}_\mathcal{L}^{(s)}$ and geometry labels $\mathbf{x}_\mathcal{O}^{(s)}$ by sampling the light parameter space around the current light estimate $\hat{\mathbf{x}}_\mathcal{L}^{(s-1)}$ and the ge-

ometry parameter space around the current geometry configuration estimate $\hat{\mathbf{x}}_{\mathcal{O}}^{(s-1)}$. We then compute the total MRF energy as

$$E^{(s)}(\mathbf{x}) = \min_{x}\{E_I(\mathbf{x}|\mathbf{x}_{\mathcal{L}}^{(s)}, \mathbf{x}_{\mathcal{O}}^{(s)})\} + E_L(\mathbf{x}_{\mathcal{L}}^{(s)}, \mathbf{x}_{\mathcal{O}}^{(s)}) + E_G(\mathbf{x}_{\mathcal{O}}^{(s)}), \tag{6.20}$$

which includes minimizing the pairwise energy $E_I(\mathbf{x}|\mathbf{x}_{\mathcal{L}}^{(s)}, \mathbf{x}_{\mathcal{O}}^{(s)})$. If the new energy $E^{(s)}(\mathbf{x})$ is lower than the previous lowest energy, we keep the proposed illumination and geometry labels $\mathbf{x}_{\mathcal{L}}^{(s)}$ and $\mathbf{x}_{\mathcal{O}}^{(s)}$, otherwise they are discarded.

As the number of geometry and illumination parameters is increasing, the choice of which dimensions of the illumination-geometry parameter domain to re-sample in order to generate proposals $\mathbf{x}_{\mathcal{L}}^{(s)}$ and $\mathbf{x}_{\mathcal{O}}^{(s)}$ becomes crucial for the effectiveness of the minimization. In our experiments we used the following proposal schedule: At some iteration s, a single light source l is chosen, and new values are generated only for the parameters of light source l and the ambient intensity to produce $\mathbf{x}_{\mathcal{L}}^{(s)}$. At iteration $s+1$, new values for the azimuth rotation and geometry class label of a single object k are generated to produce $\mathbf{x}_{\mathcal{O}}^{(s+1)}$. At iteration $s+2$ new values are generated for the 6 scalar parameters defining the 3D translation and 3D scale of a single object k to produce $\mathbf{x}_{\mathcal{O}}^{(s+2)}$. This proposal schedule is repeated every 3 iterations.

The final solution corresponds to the light parameter sample s that generated the labeling with the lowest energy:

$$\mathbf{x}^{opt} = \arg\min_{s} E^{(s)}(\mathbf{x}). \tag{6.21}$$

This method is more tolerant to local minima in the model energy (which appear often in practice) and it requires a limited number of the costly evaluations of energy $E_I(\mathbf{x}|\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}})$.

6.3.1 Proposal Generation

Sampling of the solution space to generate proposed labels is required by both minimization approached discussed above. The generation of good guesses for these proposals can significantly aid fast convergence to a good solution. In this section we discuss proposal generation.

Light directions: Proposed light source direction $\mathbf{x}_l^{dir}$ is generated by drawing a sample from a von Mises-Fisher distribution [34] with mean direction $\hat{\mathbf{x}}_l^{dir}$ and concentration parameter κ_{sample}, where $\hat{\mathbf{x}}_l^{dir}$ is the current light

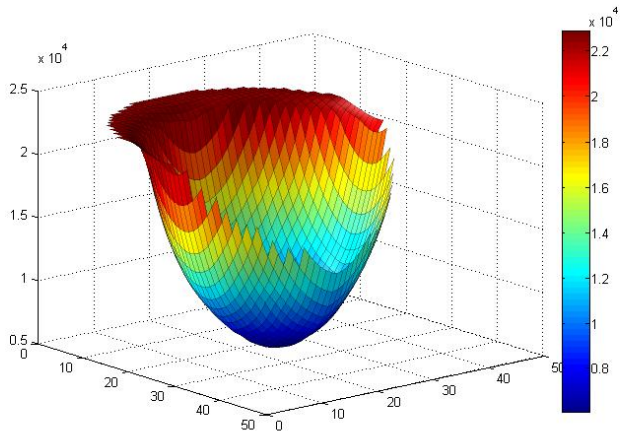

Figure 6.4: The model energy over possible directions of one light, for a simple synthetic scene.

direction estimate. The estimate from the voting algorithm is used for the first iteration. In our experiments, $\kappa_{sample} = 200$ was chosen and samples were drawn using the accept-reject algorithm.

Light intensities: The proposed intensity for light source l is computed from the current light source intensity estimate adding a random offset, drawn from a normal distribution. The same method is used for ambient intensity proposals x^{α_0}.

Pixel intensities: When using higher-order clique reduction to perform inference, proposals for the pixel labels are also required. The light proposal is kept fixed for N successive iterations, while at iteration i of the N successive iterations, pixel label i of the N different possible pixel labels is proposed for every pixel node, after which a new light parameter proposal is generated and the possible pixel labels are proposed again in the next N iterations.

Geometry parameters: The parameters used to define the geometry of an object are azimuth rotation, 3D translation, 3D scale and a geometry class label. This means that geometry for an object is defined by 7 scalars and 1 discrete value. The scalar values are drawn from normal distributions with the current value of the respective parameters used as the distribution mean. The geometry class label is drawn from a uniform distribution for each proposal.

6.4 Experimental Validation

In this section we evaluate the proposed MRF model, both quantitatively in a synthetic dataset, as well as qualitatively in real datasets. We also present results when geometry parameters are estimated simultaneously with shadows and illumination.

	 a. Exact geometry			 b. Approx. geometry			 c. Approx. geometry + noisy shadow input		
#lights:	1	2	3	1	2	3	1	2	3
Voting	7.06	6.94	8.23	5.83	11.51	13.31	20.78	28.61	29.30
NNLS [166]	3.84	6.20	6.35	13.95	15.21	14.15	33.69	32.10	33.96
MRF(HOCR [65])	3.29	5.41	8.13	5.14	14.67	13.99	14.35	20.60	22.83
MRF(2-stage minim.)	**0.44**	**1.31**	**2.36**	**2.53**	**9.06**	**8.57**	**6.97**	**12.36**	**17.77**
MRF(2-stage minim.) - w/o shadow shape prior	1.27	3.82	5.40	3.11	11.12	11.95	10.81	12.24	17.91
Number of light sources: mean error (error %)	0 (0%)	0.047 (4.7%)	0.143 (14.2%)	0 (0%)	0.309 (17.6%)	0.32 (23.8%)	0 (0%)	0.285 (26.7%)	0.33 (38%)

Table 6.1: Quantitative results on a synthetic dataset: from left to right, we show the mean error in degrees for the estimated light directions on a synthetic dataset, a) using the exact geometry to do the illumination estimation; b) using geometry approximated by bounding boxes (blue) and a ground plane; c) using approximate geometry and a noisy initial shadow estimate. For each case, we show results for scenes rendered with 1, 2 or 3 light sources. We show results obtained with the voting algorithm used for the initialization; with NNLS [166]; with our MRF model, when the MRF energy is minimized using [65]; and when the MRF energy minimized using our 2-stage approach, which achieves the best results. We also include results with our MRF model and 2-stage approach without the shadow shape-matching prior, which shows the benefits of this term. In the bottom we show the mean error in the estimated number of light sources and in what portion of images that number was estimated inaccurately.

6.4.1 Illumination Estimation

We used three different sets of images to evaluate illumination estimation results with our approach: images collected in the lab under controlled illumination conditions, real-world images of cars collected from Flickr, and the Motorbike images from Caltech 101 [110]. We overlayed a synthetic vertical pole (sun dial) onto the original images, rendered under the illumination estimated by our method, in order to visualize the results.

The weights used in our experiments were: $(w_s, w_l, w_p, w_c) = (8, 1, 1, 4)$. The upper bounds for the truncated potentials were $(t_p, t_c) = (0.5, 0.5)$. Pixel node labels were quantized to 8 values and 1000 iterations of our algorithm were performed.

Illumination estimation takes 5 to 30 minutes per image for the images in this paper, depending on image size. However, 60% to 70% of the running time is spent performing raytracing, which can be sped up significantly with a faster raytracer implementation. Table 6.2 shows the running time of our algorithm in various scenarios. Running times with voting, HOCR ([65]) and our inference approach are compared. Although HOCR is faster for 1 light source, it does not scale well as the number of light sources increases. Because of the lack of a termination criterion for our approach, we performed enough (predetermined) iterations to obtain similar or better results as with HOCR in order to be able to compare. A maximum of 200 iterations was perfomed with HOCR. With our method, the number of iterations was 200 by the number of light sources. For geometry estimation we performed 800 iterations. Experiments were perfomed on an Intel Core i7 computer with 8GB of RAM.

Synthetic Dataset

To evaluate our method quantitatively we used a set of synthetic images, rendered using a set of known area light sources. The number of light sources was randomly chosen from 1 to 3. The direction, intensity of the light sources was also chosen randomly. We examined three different cases:

Exact geometry: We used the same 3D model to render the image and to estimate illumination.

Approximate geometry: We used a 3D model that coarsely approximated the original geometry by a bounding box and a ground plane to estimate illumination.

Approximate geometry and noisy shadow input: We estimated il-

	voting	MRF (HOCR [65])	MRF (our inference method)
synthetic images (200x200 pixels, 1 light source)	2	25	33
synthetic images (200x200 pixels, 2 light sources)	2	48	61
synthetic images (200x200 pixels, 3 light sources)	2	170	95
car images (approx. 500x350 pixels)	11	414	468
car images + geometry estimation (approx. 500x350 pixels)	-	-	2036

Table 6.2: Running times (in seconds) for our algorithm, for different datasets. Times do not include shadow detection. Next to each dataset we note the average size of its images. For one light source, 200 iterations were performed with both HOCR and our MRF inference approach. For the synthetic dataset with our MRF inference method, we select the number of iterations to be performed as a multiple of the number of lights, so the runtime increases linearly with the number of light sources. In the case of HOCR [65] we did not increase the number of iterations. It should be noted that a large portion (60%-70%) of the running time is dedicated to raytracing, which could be significantly improved. The MRF inference involves considerably more time spent on raytracing than the voting initialization algorithm, since occlusions have to be computed for each light/geometry configuration proposal.

lumination parameters using a coarse 3D model, as above, and a noisy initial shadow estimate. The latter was obtained by adding random dark patches to the rendered shadow (Table 6.1.c). We used this form of noise because, on one hand our methods are relatively insensitive to spatially-uniform random noise, and on the other, in real data the errors generally affect large image regions which get mislabeled, which is emulated by this patch-based noise.

We computed the difference between the estimated light source parameters and the parameters of the true light source that was closest in direction to the estimated one. The average light source direction errors are presented in Table 6.1. We compare the results from the voting method used to obtain the initial estimate, and our MRF model. We compare the proposed inference method with a state-of-the-art method to perform inference on higher-order MRF models, the higher-order clique reduction (HOCR) technique of [65].

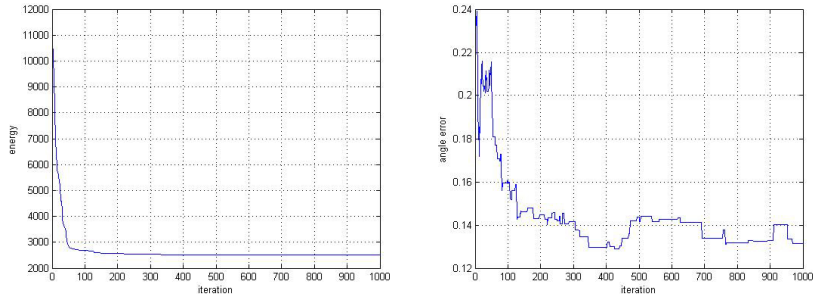

Figure 6.5: Convergence of our algorithm. Left: The energy $E(x)$ for each iteration, averaged over a set of synthetic test images (for two-stage inference, using approximate geometry and added noise to the initial shadow estimate); right: the angular error per iteration, averaged over the same test set.

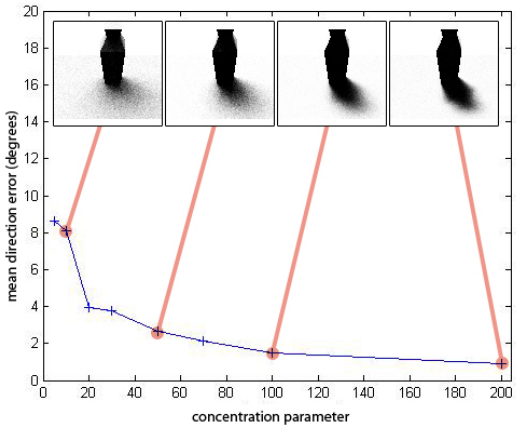

Figure 6.6: Behavior of our algorithm in the case of soft shadows. Illumination has been modeled by a vMF distribution of varying concentration κ to produce sets of images with shadows of varying "softness". Even for very "soft" shadows, the error (in degrees) in the light source direction estimate is relatively small. On the top we give examples of the images produced for sample κ values.

The results show that our method, taking advantage of the topology of this particular MRF model to efficiently perform inference, is able to achieve significantly better results, compared to our initialization method, HOCR inference on our model, as well as the non-negative least squares optimization approach of [166] (NNLS).

Furthermore, Table 6.1 shows that the shadow shape-matching prior significantly improves illumination estimates. This is more pronounced in the case of inaccurate input data, where a large number of pixels may be different between the noisy observed shadow and the one produced by the coarse geometry and true illumination. However, when there are multiple light sources, leading to a large number of potential shadow edges, the benefits of the shadow shape-matching prior are reduced.

We also evaluated the estimation of the number of light sources through our voting procedure on our synthetic dataset. Table 6.1 shows the mean error in the estimated number of light sources in that dataset. We are generally able to get a good estimate of the number of light sources. Accuracy of the number of light sources is reduced when the true number of light sources and the errors in the initial shadow estimate increase. We further evaluated our light source number estimation on the motorbike images of Caltech 101. The images we selected contained a single light source (the sun) and the average estimated number of light sources was 1.17, with the number of light sources correctly estimated 91% of the time. We should also note that any extraneous light sources identified by our voting algorithm are generally assigned low intensities during MRF inference, resulting in small errors in the synthesized cast shadows.

We further quantitatively evaluated the behavior of our method in the case of soft shadows. We rendered the set of synthetic scenes under illumination produced by a single light source modeled by a vMF distribution of varying concentration parameter κ. Lower values of κ mean a more spread-out light distribution and softer shadows. Fig.6.6 shows the error in the estimated light source direction (in degrees) as the concentration parameter of the light source changes. Even in the case of very soft shadows, our method is able to estimate the direction of illumination with good accuracy.

Real Datasets

To evaluate our approach in real images, we used the class "Motorbikes" of the Caltech 101 dataset [110] and images of cars we collected from Flickr.

In the case of "Motorbikes", we used *the same* coarse 3D model (Fig.6.11) corresponding to an average motorbike and *the same* average camera parameters for every image. In this dataset there are significant variations in geometry, pose and camera position in each individual image, deviating from our average 3D model and camera parameters. Despite these variations, our

Figure 6.7: Results for the Motorbikes class of the Caltech101 dataset. We rendered a synthetic sun dial (orange) under the estimated illumination and overlayed it on each original image. The geometry used for all instances was the same 3D model capturing an average motorbike, with the same average camera parameters.

results in Fig.6.7 show that our algorithm is robust enough to effectively estimate illumination using the same generic 3D model for all instances of a class of objects. This robustness would enable our algorithm to use results from an object detector for objects of known classes in an image, and simple common class geometry, to estimate illumination. Such an application would further require either average camera parameters or a horizon line estimator to perform illumination estimation without any input from the user.

In the case of car images collected from Flickr (Fig.6.8), the geometry was

Figure 6.8: Results with car images collected from Flickr. Top row: the original image and a synthetic sun dial rendered with the estimated illumination; Bottom row: the final shadow labels. The geometry consists of the ground plane and a single bounding box for the car.

limited even further to a bounding box approximating to the car body and a ground plane (Fig.6.11). Camera parameters were matched manually. For both Fig.6.8 and Fig.6.10 we assumed known number of light sources. Despite our initial assumption of Lambertian reflectance, the results show that our algorithm can cope with the abundance of non-lambertian surfaces in these images.

We further evaluated our algorithm in a set of images captured under controlled illumination conditions in the lab. This set includes shadows cast on a variety of textured surfaces, under 1, 2 or 3 light sources. Results on images from this dataset can be found in Fig.6.10. To estimate the illumination in this images we used rough approximate geometry, which can be seen in Fig.6.11. In Fig.6.10 we also show two synthetic examples of illumination estimation where shadows are cast on arbitrary geometry, demonstrating that we do not make any assumptions about scene geometry.

In figure 6.12 we compare the illumination estimation results between the voting algorithm we use for initialization and our MRF model. While the voting algorithm is able to find a reasonable approximation of illumination in most cases, it is usually not able to get accurate solutions in cases such as soft shadows, noise in the shadow estimate, or significant inaccuracies in the geometry. Our MRF model is more robust, especially in real-world examples; the approximate solution from the voting algorithm though offers usually a good initialization for the MRF energy minimization. Figure 6.12 compares the two for some synthetic toy examples and for real data from the Caltech101 "Motorbikes" class.

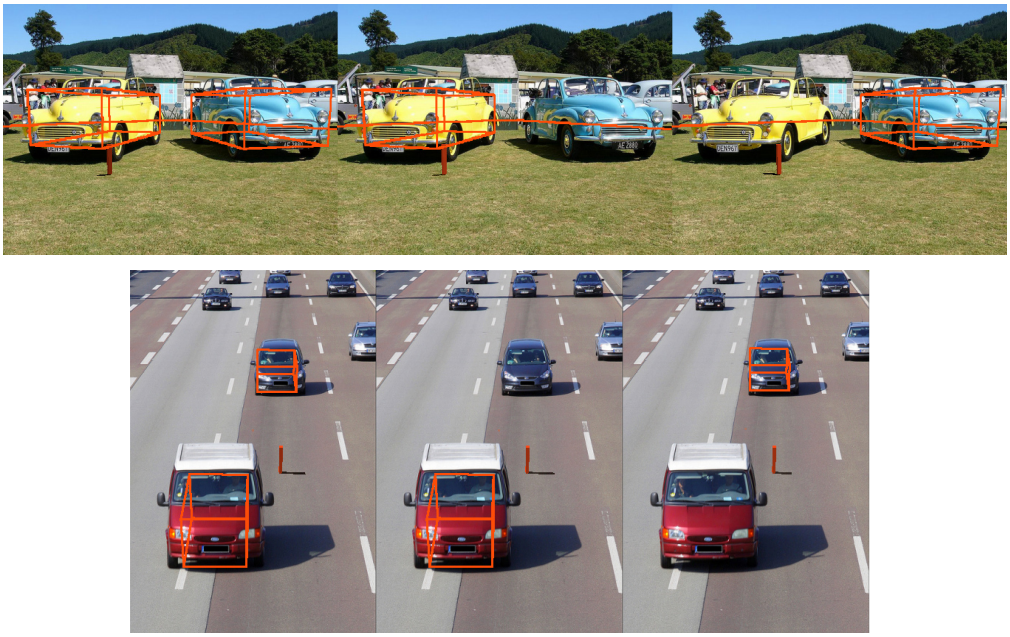

Figure 6.9: Examples of scenes with several occluders: the orange bounding boxes show the geometry provided as input to our method, and the synthetic orange sundial rendered using the estimated illumination shows our light source estimate. The illumination estimate is very stable regardless of which part of the scene we choose to model.

Failure cases

Fig.6.13 shows common cases where our algorithm fails. One general reason is challenges in shadow detection. While the shadow shape-matching prior helps our method differentiate between adjacent shadows from different occluders, it can still be challenging to correctly estimate illumination when there shadows from objects that are not modeled by the geometry are very close to or overlap shadows of interest. Furthermore, very dim shadows, as in the case of cloudy outdoor scenes, can be hard to detect, therefore not allowing us to obtain a good solution. On the other hand, coarse geometry knowledge can sometimes lead to observed shadows that cannot be explained under any illumination configuration given the coarse geometry (as in Fig.6.13.c). Inaccuracies in the placement of 3d models in the scene (e.g. with the Caltech 101 "Motorbike" images) or in the camera parameters can also lead to inaccurate illumination estimates (Fig.6.13.d).

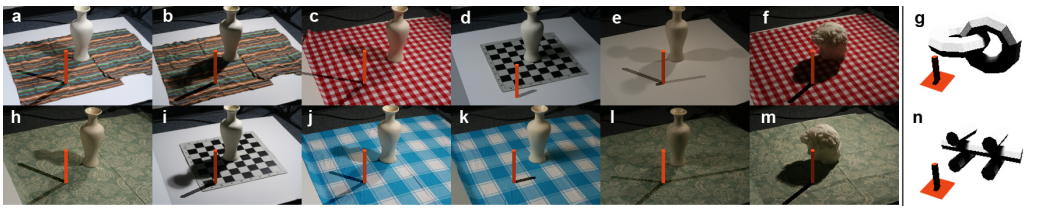

Figure 6.10: Results with captured images using different number of light sources, and different background textures. The vertical image pairs (a,h),(b,i),(c,j),(d,k),(e,l),(f,m) are captured under the same illumination. An orange synthetic sundial has been rendered under the estimated illumination and inserted into the original image. We also include a pair of results on synthetic images (g,n) that show that our method can be applied to arbitrary scene geometry, where shadows are not cast on a flat ground (mean light direction error for g,n is 2.27 degrees).

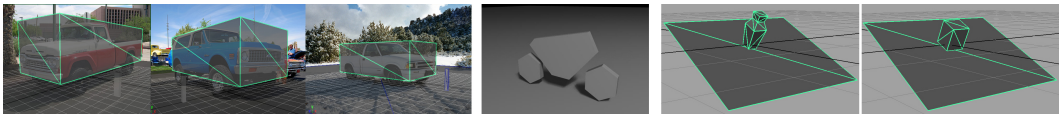

Figure 6.11: The 3D models we used to perform illumination estimation for the images of cars (Fig.6.8), motorcycles (Fig.6.8), and the images of Fig.6.10. Camera parameters where selected manually. In the case of the Caltech101 Motorcycles class, a single set of camera parameters (with orthographic projection) where used for all images. For the rest of the images, camera parameter selection was done individually for each image, although approximately.

6.4.2 Geometry Reasoning

We evaluate joint illumination and geometry/pose estimation qualitatively on the car images we collected from Flickr, as seen in Fig.6.14. The input to our algorithm in this case was the original image, a 2D bounding box around the object of interest (in this case, the car), a common ground plane, the camera parameters and a common set of 4 candidate geometric models for cars (shown in Fig.6.14). The geometric models represent 4 common car shapes. The 2D bounding box can be provided by a car detector. The camera parameters are very similar across these images, probably because of the common subject, and could approximated automatically using the information in the image EXIF

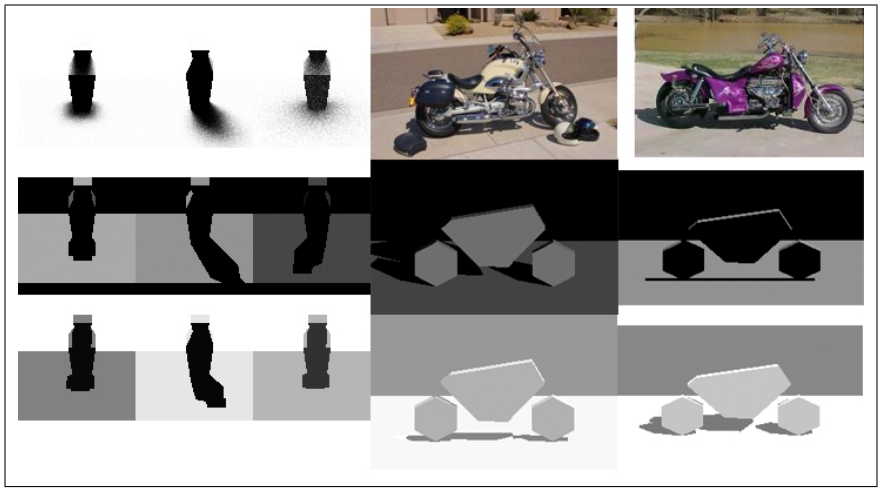

Figure 6.12: Comparison of illumination estimation results with our voting initialization algorithm and our MRF model. Top row: the original image; Middle row: the 3D model used to approximate geometry for illumination estimation, rendered with the illumination estimate obtained with the voting algorithm; Bottom row: the same 3D model rendered with the illumination estimate obtained by our MRF model, from the same input data. It is easy to see that the voting algorithm gets relatively close to the solution, but is not able to offer high accuracy especially with soft shadows or more complex images.

tag, along with horizon line estimation (and assuming the camera is at eye level of an average human). In our experiments shown in Fig.6.14 however, we set camera parameters manually.

For experiments with geometry parameter estimation we did not use our voting initialization method, because the random initial geometry reduces the benefits of such an initialization. We assumed a single light source and used a random initialization of the other parameters. A larger number of iterations (4000) was performed to obtain a solution, with larger variance for the parameter proposal generation. Despite the random initialization, our MRF model is able to obtain a satisfactory solution.

Our results show that we can approximate the orientation of the object with good accuracy (around 10 degrees), and get visually convincing estimates of scale and orientation. The object geometry is identified correctly in 3 of the 4 images below. Notice that although we could fit an infinite number of very different (and mostly incorrect) combinations of ge-

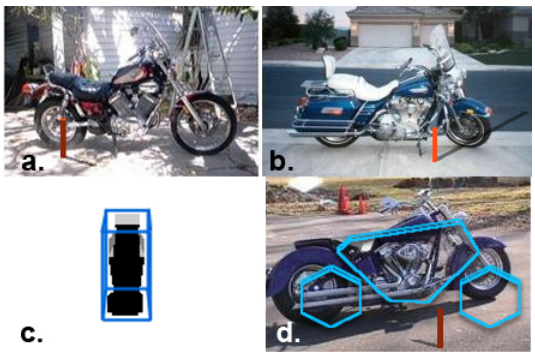

Figure 6.13: Common failure modes. Errors due to shadows (top): (a) shadows of other objects not modeled may overlap the shadows of the objects of interest, or (b) very dim shadows may not be detected, in which case our algorithm tries to use other dark image regions. Errors due to geometry (bottom): (c) approximate geometry (in blue) can have no possible way to explain observed shadows caused by the true geometry. (d) Large errors in the positioning of geometry in the scene (when geometry parameters are not estimated) affect the relative position of shadows in the image to the object geometry.

ometry/rotation/translation/scale values to the object outline obtained by GrabCut, as shown in Fig.6.14.b, the combination of the object outline and the shadow leads our algorithm to select parameter combinations close to the truth (Fig.6.14.c), while estimating the illumination at the same time. In some cases the pose estimate further improves when when combined with geometry class estimation.

An important observation is that, as the number of free parameters that define geometry grows, local minima in the energy become a bigger issue. An example of this problem is the fourth image in Fig.6.14.d, where the geometry class used for the pick-up truck corresponds to "jeep", and at the same time the size chosen for the model omits the rear part of the pick-up truck. In this case our algorithm has found a local minimum of the energy; to continue to the global minimum, a large change in scale and translation along with the change in the selected geometry class is needed. A clever selection of the dimensions which change to produce the new step on each iteration can help as the number of geometry parameters grow - for example, the geometry class could be locked to the simple bounding box for a number of iterations, expecting that the bounding box will be positioned properly over the object before we begin

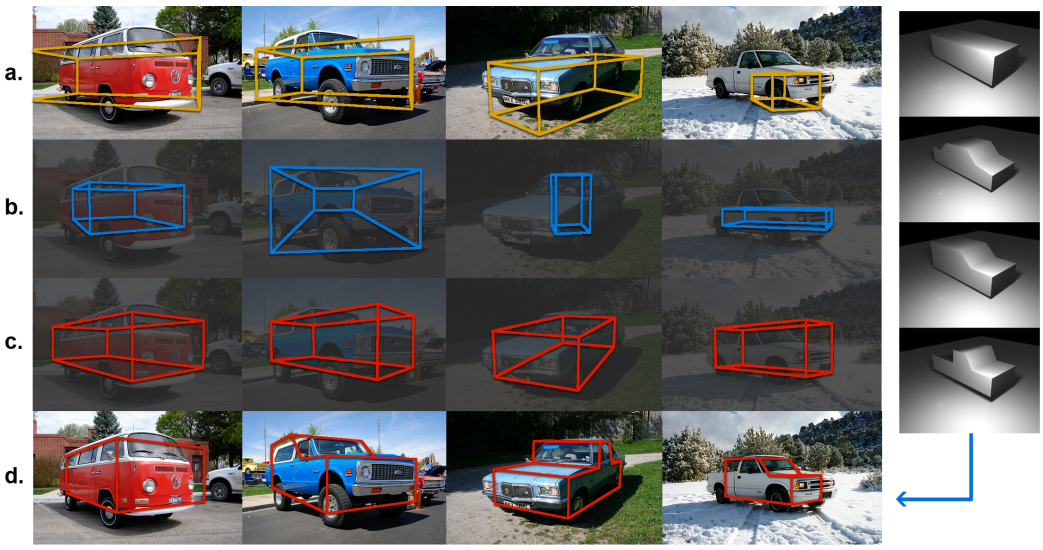

Figure 6.14: Results of joint estimation of shadows, illumination and geometry parameters. The geometry used in this case consists of a ground plane and a bounding box for the object. The geometry parameters estimated are the azimuth rotation, 3D translation and 3D scale of the object's bounding box. a) input: the original image and the initial configuration of the geometry; b) the estimated geometry when only fitting the object to the mask obtained by GrabCut; c) the geometry estimated by our method. While the object silhouette is not enough to estimate the geometry parameters, the combination of the object silhouette with information in the shadows allows us to obtain a good geometry estimate. d) Here we also allow our model to select the most probable of 4 candidate geometry classes. The estimated geometry class for each image is, from left to right: *box*, *jeep*, *sedan*, *jeep*. The 4 geometry classes are shown on the right.

examining more specific geometry classes. Random initializations of geometry very far from the true geometry can also affect the final result, but constraining the initial pose within the GrabCut mask is often sufficient.

6.5 Conclusions

In this chapter, we introduced a higher-order MRF model of illumination, which allows us to jointly estimate the illumination parameters, cast shadows and a set of geometry parameters for the occluders in a scene, given a

single image. Our model incorporates both high-level knowledge about the scene, such as illumination and geometry, and low-level image evidence. Although this leads to a complex formulation that makes inference challenging, we demonstrate that inference can be performed effectively. Our results in various datasets, demonstrate the power of our MRF illumination model. We are able to estimate the illumination parameters using the same geometry, pose and camera parameters for a large number of scenes which belong to the same class, as shown by our results on Caltech101. Bounding boxes can be sufficient approximations of occluders for our method, as is the case with our experiments with car images from Flickr. Geometry reasoning is incorporated in our model to allow estimation of the object pose in the 3D scene, as well as reasoning about the 3D geometry that best represents an object. The extensive experiments show that our approach is more general and more applicable in real-world images where other methods fail. In the future, we are interested in incorporating our method in more general scene understanding applications. Geometry parameter estimation, as presented here, is the first step towards this direction.

Chapter 7

Shape Reconstruction with a Dictionary of Shading Primitives

In the previous chapters we examined the estimation of illumination from a single image from cast shadows, which in Chapter 6 we linked with the estimation of geometry parameters. In that chapter, we linked the two inverse rendering problems of interest, illumination estimation and shape reconstruction. However, the information contained in cast shadows is not enough to obtain a good estimate of the shape of objects. As seen in the previous chapter, shadows offer only information about the object outline, which can be used to infer rough geometry or the object position and pose. Hence, in this chapter we will discuss the use of shading for the reconstruction of 3D shape.

7.1 Introduction

Shape recovery is a classic problem in computer vision and a large body of prior work exists on the subject. We examined the prior art in the area, focusing on shape-from-shading (SfS), in Chapter 2. In this chapter, we examine an approach to the problem of shape from shading based on the idea of learning shadow primitives. The goal of the work in this chapter is to infer the 3D scene structure, in the form of a normal map, from a *single* 2D image using the information contained in shading.

We capture the relationship between the appearance and geometry of image

Figure 7.1: Our method: Left, the original image [146]; center, the estimated normal map with our approach; right, a rendering of the estimated normal map under different illumination.

patches in a straight-forward way, by learning a dictionary that associates local image appearance with the underlying geometry. Each entry in the dictionary captures the geometry of a small rectangular region (*patch*) and a distribution of the possible image intensities associated with this geometry, as observed in a training set containing images of known geometry. We choose to describe the 3D scene by a normal map, containing one normal vector for each pixel to describe the orientation of the surface point which is projected to that pixel. The *input* to our algorithm is a single image, and the direction of the light source. We assume that the scene is illuminated by a single distant point light. We do not assume a specific type of surface reflectance. In our initial approach to the problem, we assume that the object surface has uniform albedo, so that an image containing only shading variations is available. Shading variations in case of variable albedo could be extracted through other methods [174].

To reconstruct the shape of a new image, we first divide the image into patches. For each image patch, we find dictionary patches that have similar appearance to the observed one. We define the distance of the image patch to the ones in the dictionary as the Mahalanobis distance between the observed appearance and the distribution of appearances that can be produced by each dictionary patch. That distribution corresponds to different parameter choices in the Ward reflectance model [188]. The dictionary patches obtained to explain an image patch constitute a set of hypotheses about the local geometry. Despite the fact that there are infinite possible geometric explanations for the

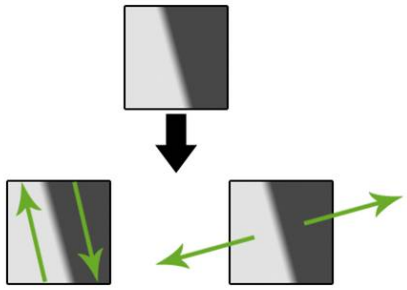

Figure 7.2: An example of the priors captured by shading primitives: The local shading pattern on the top can have any of infinite geometric explanations. The surface normals could have for example the orientations shown on the left. However, in practice such a shading pattern imposes a strong prior for the underlying geometry: it usually corresponds to an edge of a solid object, with normals oriented perpendicularly to the edge, as shown on the right. A human would also choose this as the most probable geometric explanation.

appearance of a given patch, our experiments show that certain explanations are much more probable, making our approach effective. The problem of inferring the shape of the objects in the scene becomes that of properly selecting the normal vectors given the set of local hypotheses obtained by the dictionary.

We formulate inference of the final 3D shape as a labeling problem on a Markov Random Field (MRF) model. This model allows us to chose a normal vector for each pixel that is close to the obtained local hypotheses, and at the same time satisfy anisotropic smoothness constraints. The MRF model contains one node per image pixel, with pairwise interactions between them and the node labels indicate the normal vector at each corresponding pixel, taking values in a continuous domain. We perform inference by minimizing the MRF energy using the QPBO [53, 86] and fusion-move [107] algorithms.

The main contributions of this work are the following:

1. We describe a dictionary of learned geometric primitives and the associated shading patterns. This way we learn priors to locally resolve the ambiguities inherent to the shape-from-shading problem.

2. We propose an effective way to capture the similarity between local shading patterns and learned patches using a wavelet decomposition and the

Mahalanobis distance. This allows us to handle reflectances that deviate from the Lambertian assumption.

3. We describe an MRF that combines the local geometric hypotheses to reconstruct the final normal map.

We present results in both synthetic and real results. In both cases, we demonstrate that our algorithm is able to recover both the general object shape and finer geometric details. The learned dictionaries in our experiments are trained on synthetic data, but we are able to use them to reliably reconstruct the shape of real photographs. Comparisons with other approaches [28, 177, 146] on real data show the advantages of our approach.

In the following sections we describe how image patches can be represented and how a dictionary of patches can be learned from a set of training images and their corresponding geometry (Sec.7.2), and how we can reconstruct the normal map from a test image, using the trained dictionary and formulating the problem as inference on a Markov Random Field (MRF) model (Sec.7.3). In Sec.7.4 we present results on synthetic datasets and real images with our method. Sec.7.5 concludes the chapter.

7.2 Patch dictionary

We first construct a dictionary of local geometric primitives (*patches*) from a set of training images with known geometry. Each patch in the dictionary is a small normal map of size $n \times n$, representing the local 3D geometry. Along with the geometry for each patch, we store the distribution of pixel intensities (local appearances) that can be produced by that geometry under different reflectance models, given a light source direction. We refer to each of the learned geometric primitives in the dictionary as a *dictionary patch*. By *patch appearance* we refer to the $n \times n$ grid of pixel intensities describing the appearance of an image patch or dictionary patch. By *patch geometry* we refer to the $n \times n$ grid of normal vectors representing the patch geometry.

7.2.1 Patch representation

We reduce the dimensionality of the normal map representation by applying PCA to a subset of patches from the training set and keeping the M_G first eigenvectors. Patch normal maps are therefore projected on the PCA basis and

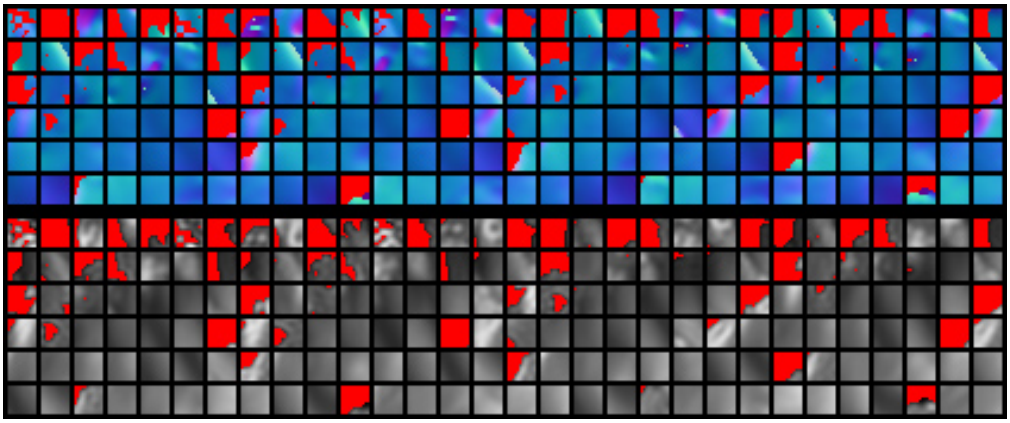

Figure 7.3: The data stored in a learned dictionary. Top: the normal map of sample dictionary patches; Bottom: the mean appearance of each dictionary patch as reconstructed from the mean of appearance wavelet coefficients. Red indicates background pixels.

represented by the M_G resulting coefficients. We choose to represent the patch appearance using a Haar wavelet basis [48]. We use Haar wavelets of order 2, using the non-standard construction, resulting in a basis of size $M_A = 16$ for appearance patches.

The distribution of appearances that can be produced by the geometry of a dictionary patch is represented by the mean and variance of the coefficients of the patch appearance. Furthermore, each dictionary patch contains a mask that indicates which pixels belong to the foreground and which (if any) to the background. Therefore, a dictionary patch $\mathcal{D}_i$ is represented by a quadruplet $\{\mathbf{G}_i, \mathbf{M}_i, \mu_i^A, \sigma_i^A\}$, where $\mathbf{G}$ are the PCA coefficients describing the patch normal map, $\mathbf{M}_i$ is the patch foreground/background mask (an $n \times n$ grid of binary values), and μ_i^A and σ_i^A are the means and variances of the coefficients of the appearances that can be produced by the patch geometry.

An example set of patch appearances and geometries from a learned dictionary is shown in Fig.7.3.

7.2.2 Dictionary construction

Let $\mathcal{T} = \left\{\left(T_k^I, T_k^G, T_k^M, \mathbf{t}_k^L\right)\right\}$ be the training set, where each training instance k consists of an image T_k^I, the corresponding normal map T_k^G, a fore-

ground/background mask T_k^M and a light source direction $\mathbf{t}_k^L$. We assume that each training instance is illuminated by a single distant light source. In order to obtain a good dictionary $\mathcal{D}$ from training set $\mathcal{T}$, we aim to learn a set of geometric primitives that could adequately describe the objects in the training set. Our approach is to: **1)** First examine only the geometry of the training set, learning a set of dictionary patches that correspond to distinct local geometric structures in our training set. **2)** As a second step, we examine the local appearance produced by each of the learned dictionary patches under different reflectances, and store statistics to describe the distribution of these appearances.

To learn the dictionary patch geometry, we first divide the geometry T_k^G of each training instance k into a set $\mathcal{P}$ of overlapping patches P_i of size $n \times n$. We then project the normal map P_k^G of each patch P_i onto the PCA basis, so that P_k^G is represented by a set of coefficients α_k^G. To decide if we should add this patch to the dictionary $\mathcal{D}$, we compute the distance between P_k and each dictionary patch $\mathcal{D}_i$ as:

$$\begin{aligned}\langle P_k, \mathcal{D}_i \rangle = &\sum_{m=1}^{M_G} \left(\alpha_k^G(m) - \alpha_i^G(m)\right)^2 + \\ &+ w_M \sum_{p=0}^{n^2} \left[P_k^M(p), \mathcal{D}_i^M(p)\right], \end{aligned} \tag{7.1}$$

where the first term is the euclidian distance of the PCA coefficients representing the geometry and the second term the difference of the foreground/background masks, weighed by a weight w_M that determines how strictly we want the foreground/background mask to match between the two patches (a large value of $w_M = 100$ was used in our experiments).

If the distance to the closest patch already in the dictionary is above a threshold θ_D, then a new dictionary patch is added to the dictionary, with the geometry and mask of patch P_k. Therefore, after all patches in the training set have been examined, a (potentially large) dictionary $\mathcal{D}$ has been constructed, containing a variety of distinct local geometric structures.

The second step is to learn the distribution of appearances that can be produced by the geometry of each dictionary patch. In order to do that, we render the normal map of each dictionary patch $\mathcal{D}_i$ using the Ward [188] reflectance model and a set $\mathcal{R}$ of different reflectance parameters, which corre-

sponds to surfaces of varying specularity, varying diffuse intensity and varying anisotropic specular properties. We project the image intensities produced by each reflectance parameter selection onto the wavelet basis, and we store the mean μ_i^A and variance σ_i^A for each appearance coefficient across all reflectance parameters.

Dictionary light source direction

Training of the dictionary assumes a known light source direction. This light source direction is used to render the local geometry under a set of different reflectances, in order to generate the distribution of appearances for each patch. We want to be able to use our approach in order to reconstruct the shape of images illuminated by arbitrary light source directions. However, it is not possible to learn a different dictionary for each possible light direction.

If we assume that we only handle Lambertian reflectances, our approach can store the normal vectors of the dictionary patches in the coordinate system of the light source. In this case, we can use the learned dictionary for images of arbitrary light source direction, simply transforming the recovered normals back from the coordinate system of the light source to that of the camera. This procedure is however not possible when non-lambertian reflectance models are accounted for.

We solve this issue by re-computing the distribution of appearances for each dictionary patch as a first step every time we are provided with a new image to reconstruct and the corresponding light source direction. Generating the distribution of appearances for a dictionary of 30000 patches, such as the one used in our experiments, takes 1-3 minutes. This time is much less than the time needed to reconstruct the image from the dictionary.

7.3 Shape reconstruction

In this section we describe how we reconstruct the geometry when provided with a new image $\mathbf{I}$ and a learned dictionary $\mathcal{D}$. We first divide the input image into a set of overlapping patches. We then find the dictionary patches in $\mathcal{D}$ that are closest in appearance to the patches extracted from the test image $\mathbf{I}$. Finally, we reconstruct the 3D shape from the results of the dictionary look-up using a Markov Random Field (MRF) model.

We divide the image $\mathbf{I}$ into a set of overlapping patches. We define an image

patch P_j for each image pixel j, so that P_j is centered at pixel j and has size $n \times n$. This way, we extract all possible image patches from the input image $\mathbf{I}$. For each image patch, we search the dictionary for dictionary patches of similar appearance. We retrieve the k_D dictionary patches that are closest in terms of appearance to image patch P_j (we define the metric to compare patch appearances in the next section, Sec.7.3.1). Because we defined image patches centered at each pixel, a given pixel i is covered by up to n^2 overlapping image patches. As a result, there are up to $k_D n^2$ dictionary matches that include pixel i, with each dictionary match defining a normal vector for pixel i. Each of these results is considered a hypothesis about the vector at pixel i.

Because of the dependency of patches on scale, we repeat this search for a set of different scales $\mathcal{S}$. We use re-scaled versions of the original image, at scales both coarser and finer. We examine every patch at the coarsest scale. At finer scales, we only examine those image patches that have image variance above a given threshold (0.001 in our experiments). Moving to finer scales, the patches get smaller relative to the image. As a result, the average image variance per patch reduces, so that only finer details are examined at finer scales (see Fig.7.4). The dictionary matches of size $n \times n$ at each scale are then re-scaled to the scale of the original image. As a result, the final set of dictionary matches contains patches of varying sizes, corresponding to the different image scales used for the search.

The above procedure generates up to $|\mathcal{S}|k_D n^2$ normal vector hypotheses for each image pixel i. From this large set of hypotheses, we keep only the k normal vectors that correspond to the k dictionary patches with the lowest matching cost that contain this image pixel. These candidate normal vectors will be subsequently used in the MRF optimization described in section 7.3.2 to obtain the final normal map.

7.3.1 Dictionary search

To determine how well a dictionary patch (consisting of a normal map patch and a set of appearance statistics) matches an image patch (consisting of a patch of image intensities) we use the Mahalanobis distance.

Let P_j be an image patch consisting of appearance P_j^A (a $n \times n$ patch of per-pixel intensities) and a foreground/background mask P_j^M. Projecting the foreground pixels of appearance P_j^A onto the appearance wavelet basis, we obtain a set of coefficients α_j^A that describe the image patch appearance. We compute the distance between the appearance of P_j and that of a dictionary

patch $\mathcal{D}_i$ by the Mahalanobis distance:

$$D_A(\mathcal{D}_i, P_j) = \sqrt{\sum_{m=1}^{M_A} \frac{\left(\alpha_j^A(m) - \mu_i^A(m)\right)^2}{\left(\sigma_i^A(m)\right)^2}}, \tag{7.2}$$

where μ_i^A and σ_i^A are the mean and variance of the appearance coefficients of the appearances produced by dictionary patch $\mathcal{D}_i$ under different reflectances, as computed during training [1].

To compute the quality of the match between dictionary patch $\mathcal{D}_i$ and image patch P_j, we also compute the similarity of the foreground/background masks of the two patches:

$$D_M(\mathcal{D}_i, P_j) = \frac{1}{n^2} \sum_{x=1}^{n} \sum_{y=1}^{n} \left[\mathcal{D}_i^M(x,y) = P_j^M(x,y)\right], \tag{7.3}$$

where $\left[\mathcal{D}_i^M(x,y) = P_j^M(x,y)\right] = 1$ if both masks agree for pixel (x,y) and 0 otherwise.

The final cost of using dictionary patch $\mathcal{D}_i$ to explain image patch P_j is then:

$$cost(\mathcal{D}_i, P_j) = D_A(\mathcal{D}_i, P_j) + w_M D_M(\mathcal{D}_i, P_j), \tag{7.4}$$

where w_M is a weight that determines how strictly we want the foreground/background mask to match between the two patches (the same as in Eq.7.1). Therefore, the best matches in the dictionary to explain an image patch P_j will be able to produce similar appearances while at the same time will have a similar foreground/background mask.

7.3.2 Combination of dictionary matches

Having obtained a set of dictionary matches, we then produce one normal map per scale that contains the average of the best k matches from the dictionary for each pixel. The normal maps for each scale are combined to produce a first guess about the final normal map, by setting the normal of each pixel to be the average normal of the finest scale that has been recovered for that pixel. The results we obtain at each scale and their combination to produce

[1] We have assumed that covariances between appearance coefficients are 0, which lead to no significant deterioration in results, but significantly faster training and testing.

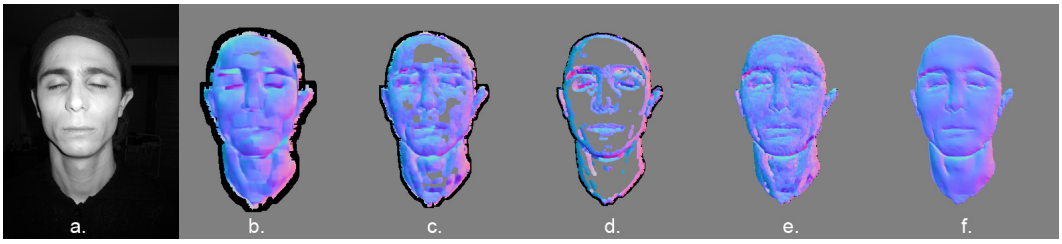

Figure 7.4: Combining matches over different scales to produce an initial guess about the normal map. a) original image; b-d) the normal maps produced by averaging dictionary matches at 3 different scales; e) the combination of all scales to produce an initial guess about the normal map; f) the final result from our method.

the initial guess are shown in Fig.7.4.

We refine this initial guess to produce the final normal map by modeling the problem as an MRF model. Through the MRF optimization, we estimate a normal map for the image that is both close to the discovered dictionary matches and that satisfies anisotropic smoothness constraints.

Our MRF model can be represented by a 4-connected 2D lattice, where each node corresponds to an image pixel. Each random variable x_i at pixel i indicates a normal vector $\mathbf{n}_i$. Therefore, the labels x_i take values from a continuous domain. The energy of the MRF model is:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{I}} \phi_i(x_i) + w_2 \sum_{i,j \in \mathcal{N}} \psi_{ij}(x_i, x_j), \tag{7.5}$$

where $\mathcal{I}$ is the set of image pixels, $\mathcal{N}$ is the set of neighboring pixels in the 4-connected grid, $\phi_i(x_i)$ is the singleton potential that associates the labels x_i with the geometry hypotheses recovered from the dictionary $\mathcal{D}$ and $\psi_{ij}(x_i, x_j)$ is the pairwise potential associating neighboring pixels i and j. The w_2 was set to 0.1 in our experiments.

The form of the *singleton potential* is:

$$\phi_i(x_i) = w_i^I \min_j \left\{ \frac{\arccos\left(\mathbf{n}(x_i) \cdot \mathbf{n}(\mathcal{D}_j)\right)}{cost(\mathcal{D}_j)} \right\}, \tag{7.6}$$

where $\mathbf{n}(x_i)$ is the normal vector at pixel i indicated by label x_i, $\mathbf{n}(\mathcal{D}_j)$ is the normal vector at pixel i as predicted by match j, and $cost(\mathcal{D}_j)$ is the cost associated with match $\mathcal{D}_j$. Furthermore, w_i^I is a weight that corresponds to

how reliable we expect the dictionary matches at pixel i to be.

We express w_i^I based on two observations: dictionary matches are more reliable when there is enough local image variability (flat image regions are the least informative), and dictionary matches are not reliable when the matches in different scales differ significantly from each other. Therefore, we define w_i^I as:

$$w_i^I = \frac{\sigma_i}{1 + q(i)}, \tag{7.7}$$

where σ_i is the local image variance at pixel i, which is computed as the variance of the image pixel intensities in a 6×6 patch centered at pixel i. The term $q(i)$ represents how much the recovered dictionary patches differ at pixel i, and is defined as:

$$q(i) = \frac{1}{\pi} \sum_{s=0}^{|\mathcal{S}|} \sum_{j} \arccos\left(\mathbf{n}(\mathcal{D}_j^s) \cdot \overline{\mathbf{n}}_i\right), \tag{7.8}$$

where $\mathcal{S}$ is the set of different scales we are examining, $\mathcal{D}_j^s$ indicates the j-th recovered dictionary patch for pixel i using scale s, and $\overline{\mathbf{n}}_i$ is the normal vector at pixel i obtained by averaging the normals at pixel i from all recovered dictionary matches at all scales.

The *pairwise potentials* $\psi_{ij}(x_i, x_j)$ enforce smoothness between the normals of neighboring pixels i and j:

$$\psi_{ij}(x_i, x_j) = w_{ij} \arccos\left(\mathbf{n}(x_i) \cdot \mathbf{n}(x_j)\right), \tag{7.9}$$

where w_{ij} is a weight computed as a function of the image gradient between pixels i and j:

$$w_{ij} = \max\left\{0, 1 - w_\nabla \nabla I_{ij}\right\}, \tag{7.10}$$

and w_∇ determines how sensitive the smoothing term is to image gradients (we set $w_\nabla = 4$ in our experiments).

An *alternative formulation* ψ_{ij}^L for the pairwise potentials could take into account the image intensities to increase the detail in the final result. This however necessitates the assumptions of a reflectance model. To produce this alternative pairwise potential form, we assume Lambertian reflectance. We introduce the image intensities in our model using the differences of intensities

between neighboring pixels i and j:

$$\psi_{ij}^{L}(x_i, x_j) = \psi_{ij}^{L}(x_i, x_j) + \\ + w_L \left((I_i - I_j) - (\max\{0, \mathbf{n}(x_i) \cdot \mathbf{d}\} - \max\{0, \mathbf{n}(x_j) \cdot \mathbf{d}\}) \right)^2, \tag{7.11}$$

where $\mathbf{d}$ is the light direction and I_i and I_j are the image intensities at pixels i and j. The weight w_L modulates the contribution of the new Lambertian term to the final solution. In practice, this modified form of the pairwise potentials increases the detail of the result, but it also introduces artifacts and shape distortions, especially in the case of real photographs with non-lambertian surfaces. Therefore, unless for a certain application it is safe to assume Lambertian reflectance, this alternative formulation of the pairwise potentials would not be recommended.

We infer the final normal map by minimizing the MRF energy over the labels $\mathbf{x}$:

$$\mathbf{x}^{opt} = \arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{7.12}$$

We chose to use the QPBO [53, 86] and fusion-move [107] algorithms to perform inference. The QPBO algorithm is used to solve a binary MRF labeling problem between the current set of node labels $\hat{\mathbf{x}}$ and a set of proposed labels $\mathbf{x}'$. The solution is initialized to our initial guess about the normal map, produced by keeping the average normal of the finest scale available for each pixel. We perform a predefined number of iterations, and at each iteration we generate the set of proposed normals (indicated by labels $\mathbf{x}'$) by adding a small random offset to each normal vector in the current solution $\hat{\mathbf{x}}$.

7.4 Experimental Evaluation

We evaluated our method on both real (Fig.7.6) and synthetic (Fig.7.7) data. For evaluation on synthetic data, we used a set of 3D models rendered assuming Lambertian reflectance. The set consisted of 6 models of real objects captured with a 3D scanner [178, 23] and rendered from 142 different viewpoints and a set of 2.5D range images of 11 different objects [58], captured from 66 different viewpoints. We used a subset of the viewpoints available, resulting in a set of 150 images. We used leave-one-out cross-validation to evaluate our algorithm: we reconstructed the shape from an image of model i using a dictionary trained on all models other than i (excluding multiple views of the same object as well).

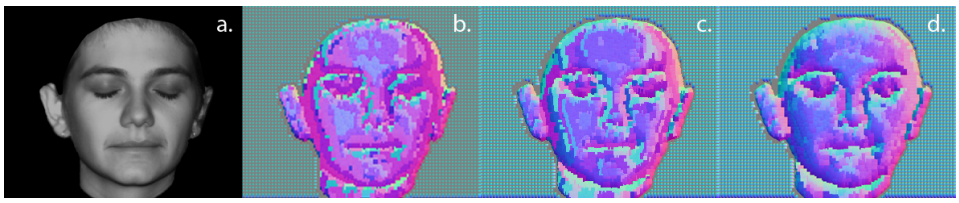

Figure 7.5: The effect of patch size demonstrated using a special dictionary of spherical patches: a) the original image; b-d) the normal maps reconstructed from the original image using a dictionary that contains only patches representing spherical surfaces. We reconstruct the original image using a dictionary of 4×4 pixel patches (b), an image enlarged by a factor of 2 using 8×8 patches (c), and an image enlarged 4 times using 16×16 pixel patches. It can be seen that larger patch sizes offer a much better support to infer local curvature. No detail was added while enlarging images.

We used 4 scales (1/4, 1, 2 and 4 times the size of the original image) to recover matching patches from the dictionary. The smaller scale better captures the overall shape of the object, while finer scales can better capture detail. A total of 5000 iterations was performed during MRF inference. The running time of our algorithm was 20-40 minutes per image, depending on image size and the size of the dictionary (running time measured on an Intel Core i5 machine). Training for a dataset of 150 images takes slightly over an 1hr. We integrated the normal maps estimated by our method using the M-estimator [1], in order to produce the final 3D surfaces (Fig 7.8).

For our experiments, we used a dictionary of 30000 patches of size 12×12 pixels. We used a Haar wavelet basis of size 16 and the first 90 PCA eigenvectors for the patch normal maps. We observed that dictionaries of at least 10000 patches were necessary in order to get satisfactory reconstructions, while having more than 30000 patches (for the selected patch size) was usually only marginally beneficial to our results.

Furthermore, it was apparent from our experiments that the patch size needs to be at least 8×8 pixels in order to reasonably capture local shape, with larger patch sizes offering better results. We can demonstrate this through a custom dictionary containing only patches of *spherical surfaces*. Reconstructing an image from that dictionary corresponds to assuming that the surface is locally spherical. We examined different patch sizes, rescaling the image so that the relative size of the patch to the image remains constant. For exam-

ple, the image we used for patch size 16×16 was enlarged 4 times compared to the original, used for patch size 4×4 (no details where added by enlarging the images). The reconstructed 3D shape obtained is significantly more accurate with patch sizes larger than 8×8 pixels, with the best results obtained by the largest patch size we tried, 16×16 pixels (see Fig.7.5). This dictionary, constructed by spherical surfaces, ignores the high-frequency information, capturing only the overall curvature of each patch. Therefore, our result shows that relatively large patch sizes are required to reliably capture the local curvature of surfaces. However, larger patch sizes require constructing larger dictionaries. In our experiments, the choice of a 12×12 pixel size was sufficient.

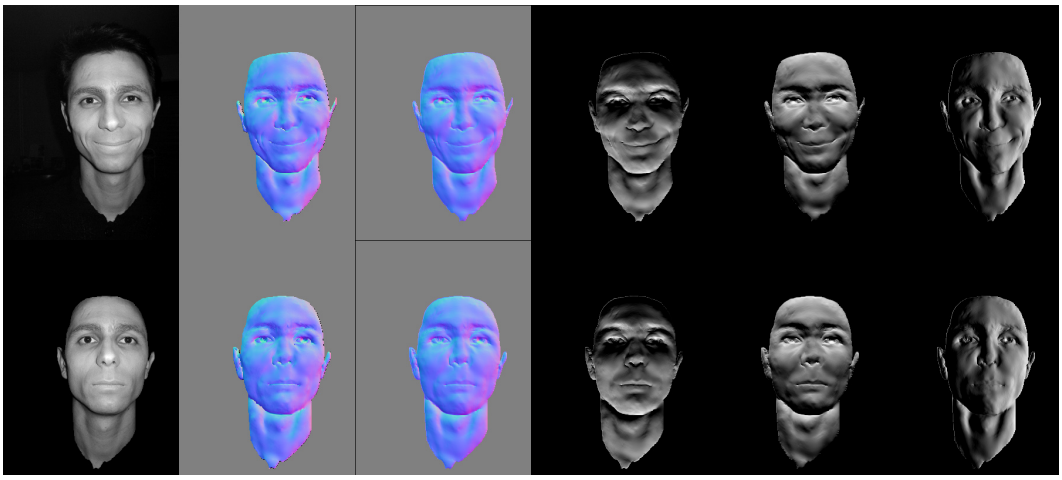

Figure 7.6: Reconstruction from a real photograph. From left to right, original image (from [146]); the normal map estimated with our method; the normal map after integrating our estimate using the M-estimator [1]; 3 rendered images with the normal map we estimated and different light directions.

In our experiments, our method significantly outperforms previous shape-from-shading approaches (Fig.7.9). It is able to reliably capture the general orientation of surfaces and is able to reconstruct much more local detail than other approaches [28, 177, 146]. This can be attributed to the fact that most shape-from-shading approaches rely on some kind of smoothness constraint, whereas in our case such constraints are replaced by the learned primitives. Smoothness needs to be enforced much more weakly during our MRF inference, allowing the solution to retain a lot of local detail. In our experiments with real

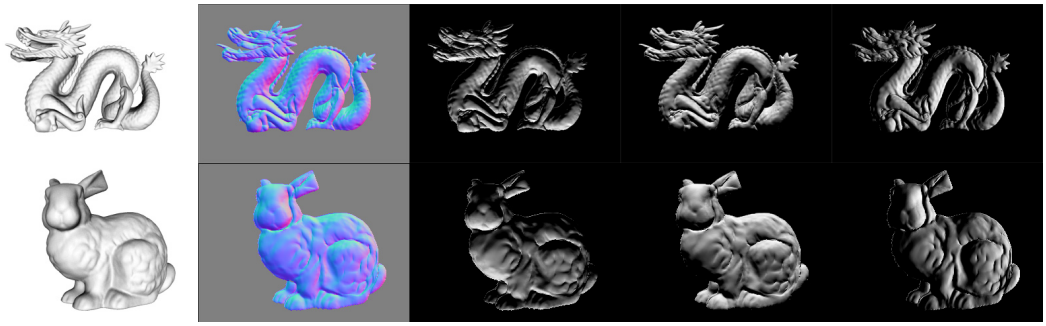

Figure 7.7: Reconstruction of normal maps of synthetic images. The images are generated by rendering depth maps of objects collected by 3D scanning [178, 23]. We show the reconstructed normal maps and renderings of the reconstructed shape under different illuminations.

data, our method also outperforms the shape-from-shading approach of [146] that applies to specific cases of the problem that can be well-posed. The ability of our method to handle surfaces that are not lambertian is one extra reason for the improved performance on real images. The results in Fig.7.11 show that the shape reconstruction with our approach is not significantly affected by surface reflectances that deviate from the Lambertian model. Fig.7.11 further shows that our method can handle reflectance parameters that are not included in the set of reflectance parameters used during training. The use of the Mahalanobis distance further allows us to cope with images that are not photometrically calibrated (e.g. underexposed images), which can be challenging when matching the local patch appearance, since in the set of reflectances used to build the distributions of appearances in the dictionary we have also included surfaces with lower uniform albedo.

In Fig.7.10 we show the reconstructed 3D shape with our method for Mozart, when illuminated from 3 different light directions.

One weakness of our method is that the quality of the results diminishes in the case of objects with large flat surfaces, indicating that flat patches are significantly more ambiguous than patches that contain even slight shading variations.

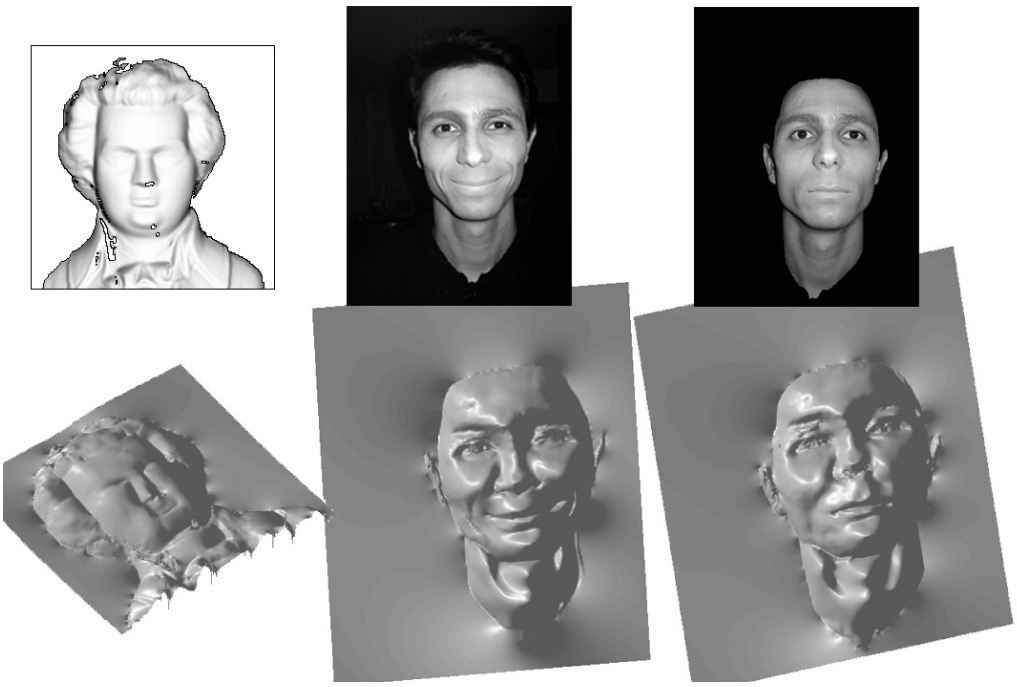

Figure 7.8: Examples of 3D surfaces reconstructed from the normal maps estimated with our method, using the M-estimator [1].

7.4.1 Image relighting

In Fig.7.12 we show results with image relighting of real objects. The input to our method is the original image, a foreground/background mask and the light direction. We extract the brightness as the maximum of the RGB color channels for each pixel. We apply our method to estimate the normal map. We finally render the estimated normal map using Lambertian reflectance under a new light direction. The final image is formed by transferring the hue and saturation from the input image, and using the brightness from our rendering. The result is realistically relighted images, as seen in Fig.7.12.

7.4.2 Refining coarse geometry

In this section we demonstrate how we can refine a coarse normal map using our approach. The initial geometry is collected using a Microsoft Kinect (a consumer device that includes a 3D scanner and a camera). The collected

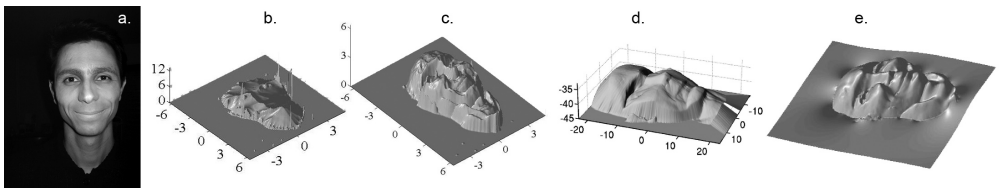

Figure 7.9: Comparison of our method with other approaches: a) original image; Surface estimates by: b) [28]; c) [177]; d) [146]; e) our approach. Please note that our approach captures both the overall shape of the object as well as the details better, resulting in a 3D face with clearly discernible features and a closer resemblance to the original.

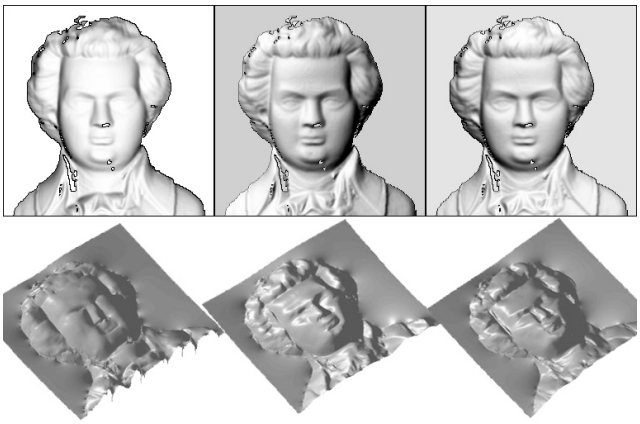

Figure 7.10: Shape reconstruction with different light directions with our method: Mozart, illuminated from 3 different light directions.

data are an image and a depth map. The depth values in the depth map are reliable but of low resolution. Therefore, computing the normal vectors from the depth map leads to unsatisfactory results, even when smoothing is used on the depth values, as shown in Fig.7.13. Furthermore, the collected depth map contains a lot of wholes, especially around the occlusion borders of objects. We can use our approach to refine such results.

To refine coarse known 3D geometry, we make a very simple modification to our dictionary search cost $cost(\mathcal{D}_i, P_j)$ in Equation 7.4: We add a matching cost term between the normal map of dictionary patch $\mathcal{D}_i$ and the coarse (smoothed) normal map that is known for the test patch P_j. The new cost

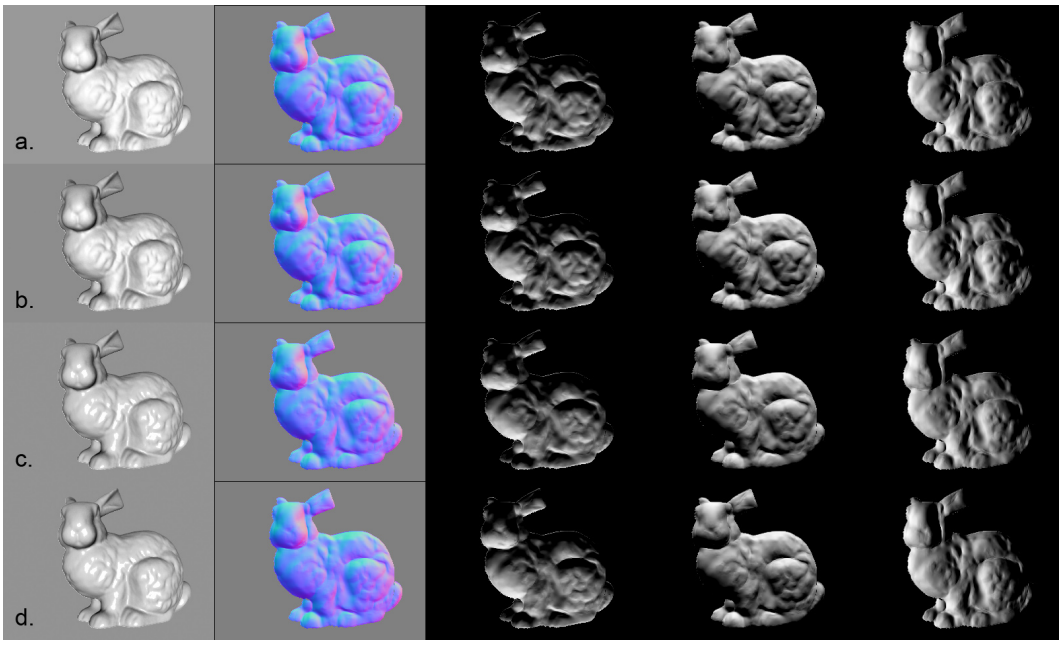

Figure 7.11: Behavior of our approach for different surface reflectances. A synthetic image has been rendered using the Ward reflectance model and different reflectance parameters, corresponding to surfaces of varying specularity. From left to right, the input image, the normal map estimated with our method, and three renderings of the normal map under novel light directions. Rows a-c show results with three different reflectance parameter choices of increasing specularity. The results show that our approach is not significantly affected by reflectances that deviate from the Lambertian model. In (c), the reflectance of the object has a stronger specular component that the most specular component used to generate the local appearance distribution for each dictionary patch while training. For comparison, in (d) we reconstruct the same image as (c), but with a dictionary trained using reflectance parameters that include surfaces as specular as or more specular than the test image. It can be seen that the results in (c) are comparable to the results in (d), demonstrating that the choice of reflectance parameters while training is not crucial for the results.

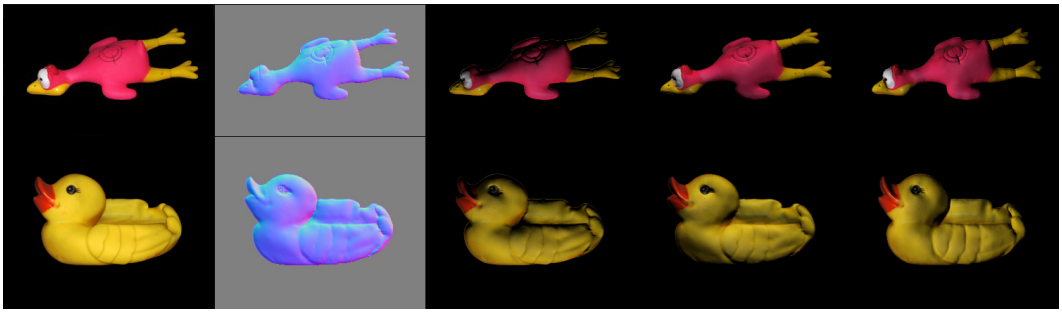

Figure 7.12: Two real objects, including some albedo variations, relighted with our approach. First column: the input image; second column: the estimated normal map with our method; Last three columns: three relighted versions of the original image.

becomes:

$$cost'(\mathcal{D}_i, P_j) = cost(\mathcal{D}_i, P_j) + w_G \sum_{m=1}^{M} \left(\alpha_i^G(m) - \alpha_j^G(m)\right)^2, \tag{7.13}$$

where $\alpha_i^G(m)$ is the m-th coefficient of the geometry of dictionary patch $\mathcal{D}_j$, $\alpha_j^G(m)$ is the m-th coefficient of the *coarse* geometry of the test patch j, and w_G is a weight (we set that weight to 1 for our experiments).

With this modification, the resulting dictionary matches that try to explain patch P_j not only match the appearance of that patch, but also the known coarse geometry. Because we are interested to refine the geometry, and therefore we assume the known geometry is coarse, we select a small value for M in Eq.7.13. Thus, we require only the first few PCA coefficients of the geometry to match, corresponding to the low-frequency geometric information. The geometric details are constrained only by the shading in this matching. We chose $M = 6$ in our experiments shown in Fig.7.13.

Figure 7.13 shows the results for an example scene captured using a Kinect. Our method is able to complete the holes in the collected depth map, and to obtain a convincing normal map. We show the normal maps we obtain from the Kinect depth data using various levels of smoothing on the depth values for comparison.

The geometry refinement results we present here show how the approach we presented in this chapter could be combined with the illumination and coarse

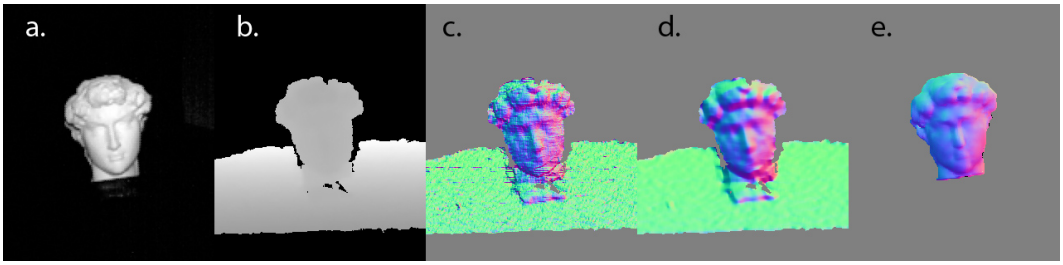

Figure 7.13: Refinement of geometry captured with a Kinect: a) the image captured by the Kinect; b) the depth map captured by the Kinect (notice that there are gaps around the edges of the object); c) the normals computed by the depth map; d) normals computed by the depth map after gaussian smoothing of the depth values; e) the normals computed by refining the smoothed normal map (d) using our method. Notice that we have correctly completed all the object edges, as well as increased the detail in the object while removing noise.

geometry estimation of Chapter 6 in order to provide a complete treatment to the problem. The missing link for this integration is the separation of albedo from shading. While the work in intrinsic images offers such capabilities, the quality of the results is not sufficient for the purpose of reconstructing shape in complex natural images. However, the concept of dictionaries of shading primitives could be extended to include factorizations of local appearance into albedo and shading components. The shading component would then be constrained by the requirement to correspond to a plausible geometric surface, making the extraction of intrinsic images significantly more robust. We believe that this would be a very interesting direction for future research.

7.5 Conclusions

In this paper we presented a data-driven approach to the problem of shape-from-shading from a single image. We described how we can build a dictionary that captures in a straight-forward way the correlations between different structures in local shading and geometry. Such a dictionary can provide a set of local hypotheses to explain the geometry underlying an observed image. The final surface structure can be recovered by combining these hypotheses in an MRF model. The advantages of a data-driven approach are that it removes a lot of typical considerations in SfS algorithms, such as boundary conditions

or the choice of camera model, and enables us to explicitly deal with surfaces that deviate from the lambertian reflectance model. The results with this approach outperform previous shape-from-shading approaches, even when such approaches make significantly more assumptions than ours. Futhermore, the work presented in this chapter opens several areas for future research:

- We can try to learn a dictionary for the more complex case of global illumination, in order to model interreflections. Apart from using a training set rendered under global illumination, we should choose the frequencies of patch appearance in an informed way [50] to obtain good matches from the dictionary. The expression of appearance on a wavelet basis facilitates the development of the appropriate distance measures.

- A second important extension to this work would be the incorporation of albedo variations in the learned dictionary. Specifically, a dictionary that contains possible decompositions of local appearance to shading and albedo patterns could be learned. This approach resembles prior work in intrinsic images [174] - the use of such a decomposition in a shape reconstruction framework would offer however a powerful constraint: the shading component of the decomposition must correspond to a plausible 3D surface. Furthermore, the decomposition of local appearance on a wavelet domain (or similar) allows the association of specific spatial frequencies to albedo or shading, facilitating such decompositions.

- As seen from the previous two areas for future work, one of the most important components of such approaches is the choice of a basis and distance metric for local appearance. Our choices enabled the use of this method on real photographs that are not photometrically calibrated, and where surfaces deviate from the Lambertian assumption. There is, however, potential for significant improvements given more research on how appearance distance metrics affect the reconstructions.

- An example of possible future advances related to the way appearance is defined would be the estimation of reflectance parameters. The current approach models appearance as a distribution produced by a set of difference reflectance parameters. Instead of assuming reflectance is an unknown parameter and effectively treating it as noise, one could define the distribution of appearances for a given geometric primitive as a high-dimensional manifold over reflectance parameters, based on a

low-parametric BRDF model (e.g. [121]). Such modeling could enable the estimation of reflectance jointly with shape reconstruction.

- The distribution of appearances for each geometric primitive is re-computed for the light direction of each input image that needs to be reconstructed. This naturally allows the application of our method in cases of complex natural illumination. In such cases, though, the matching of observed shading patterns to dictionary patches can become challenging.

The future work described here could potentially allow for a complete, data-driven treatment of the shape-from-shading problems in general images. Even if such a complete treatment of this problem is not possible, given the inherent limitations of shape-from-shading, our approach allows for a useful amount of information to be extracted from shading in a large class of images. It therefore can allow the combination of shading with other sources of information in more complex multi-cue frameworks, towards the ultimate goal of image understanding.

Chapter 8

Conclusions

In this thesis we examined two of the three inverse rendering problems, those of estimating illumination and reconstructing 3D shape from shading variations. We presented approaches that attempt to solve these problems using probabilistic approaches that relax the strong assumptions of prior work.

In illumination estimation, we relaxed the onerous assumption of accurate knowledge of geometry, presenting three different techniques that estimate illumination from the cast shadows in an image: 1) a method based on EM and the modeling of illumination as a mixture of von Mises-Fisher distributions, which is able to model soft shadows as well; 2) a method that estimates illumination by associating the light source parameters with the observed shadow edges in the image, which offers competitive performance and low computational complexity; 3) a method that combines ideas from the two previous ones and models the creation of cast shadows as a Markov Random Field model. This approach not only offers a robust and effective method for the estimation of illumination in real images, but it also provides a powerful formalism that can incorporate other facets of the problem in a unified framework. Along these lines, we introduce parametrizations of geometry into this framework and demonstrate that it is possible, through this MRF formulation, to jointly estimate all major components of the problem at the same time: cast shadows, illumination and geometry parameters.

Although through this approach we are able to infer geometry parameters that define the rough 3D geometry of occluders in the scene, the information contained in cast shadows is not adequate to estimate detailed 3D shape. Moving towards this direction, we proposed an approach that infers 3D shape from the local shading variations in the observed surfaces. This approach

is based on the idea of building a dictionary of geometric primitives, and learning the relationship between the local geometry and appearance. When reconstructing an image, the hypotheses about local 3D geometry produced by this dictionary are combined to infer the final 3D surface through an MRF model. This approach is demonstrated to offer a reliable way to overcome the challenges and ambiguities that plague the shape-from-shading problem. We are able to estimate the 3D shape of real objects, even when they exhibit non-lambertian reflectance, and obtain results that are superior to prior work while at the same time relaxing many of the assumptions that work relied on.

The work in both areas aims to make the application of inverse rendering problems viable in real-world scenarios where we examine natural scenes that may be of significant complexity, and the extra information provided, apart from a single image, is limited and unreliable. The result is that the methods we propose are useful not only for direct practical applications, such as inserting a synthetic object in a photograph or relighting a photograph obtain with flash illumination, but also as components in the ultimate goal of scene understanding. On one hand, we formulated our methods in modular frameworks based on graphical models. Therefore, new components can be added to our framework, or the entire framework can be used as a component in a larger graphical model that captures the interaction of various sources of information about a scene, such as object detectors. On the other hand, we significantly relax the assumptions about the initial knowledge of the scene. Hence, our methods can utilize approximate information obtained with automatic methods, which would be a necessity in order to integrate them in larger scene understanding frameworks.

This integration towards larger scene understanding tasks would be one major direction for future research, based on the work presented in this thesis. A second general line of work would be the extension of the ideas we presented about 3D shape reconstruction, not only to improve performance, but: 1) to allow the more concrete integration of reflectance parameters in the problem, and 2) to incorporate a decomposition of local appearance into albedo and shading. The latter would introduce ideas similar to those proposed for the extraction of intrinsic images to our approach, but would combine them with powerful geometric constraints towards a complete treatment of the shape-from-shading problem in the case of general images with varying albedo.

Bibliography

[1] Amit Agrawal and Ramesh Raskar. What is the range of surface reconstructions from a gradient field. In *ECCV*, pages 578–591. Springer, 2006.

[2] Asem M. Ali, Aly A. Farag, and Georgy L. Gimel'Farb. Optimizing binary mrfs with higher order cliques. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 98–111, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] Sanjay M. Bakshi. Shape from shading for non-lambertian surfaces. In *ICIP*, volume 94, pages 130–134, 1994.

[4] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

[5] Stephen T. Barnard. Stochastic stereo matching over scale. *International Journal of Computer Vision*, 3:17–32, 1989. 10.1007/BF00054836.

[6] Jonathan T Barron and Jitendra Malik. High-frequency shape and albedo from shading using natural image statistics. In *CVPR*, 2011.

[7] Harry G. Barrow and J. M. Tenenbaum. Retrospective on "Interpreting line drawings as three-dimensional surfaces". *Artificial Intelligence*, 59(1-2):71–80, February 1993.

[8] Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:383–390, 2003.

[9] Peter N. Belhumeur and David J. Kriegman. What is the set of images of an object under all possible lighting conditions. *IJCV*, 28:270–277, 1998.

[10] Peter N. Belhumeur, David J. Kriegman, and Alan L. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999.

[11] J. Ben-Arie and D. Nandy. A neural network approach for reconstructing surface shape from shading. *Image Processing, International Conference on*, 2:972, 1998.

[12] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, B-48:259–302, 1986.

[13] M. Bichsel and A. P. Pentland. A simple algorithm for shape from shading. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–465, 1992.

[14] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[15] Endre Boros and Peter L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 123(1-3):155–225, 2002.

[16] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:359–374, 2001.

[17] Yuri Boykov and Gareth F. Lea. Graph cuts and efficient n-d image segmentation. *IJCV*, 70(2):109–131, November 2006.

[18] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.

[19] Matthieu Bray, Pushmeet Kohli, and Philip H. S. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *ECCV*, pages 642–655, 2006.

[20] Michael J. Brooks. Two results concerning ambiguity in shape from shading. In *AAAI*, pages 36–39, 1983.

[21] Michael J. Brooks and Berthold K. P. Horn. Shape from shading. chapter Shape and source from shading, pages 53–68. MIT Press, Cambridge, MA, USA, 1989.

[22] Frederic Courteille, Alain Crouzil, Jean denis Durou, and Pierre Gurdjos. Towards shape from shading under realistic photographic conditions. In *Proceedings of the 17th International Conference on Pattern Recognition - ICPR 2004, vol.2*, pages 277–280, 2004.

[23] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, SIGGRAPH '96, pages 303–312, New York, NY, USA, 1996. ACM.

[24] B. T. Porteous D. M. Greig and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51:271–279, 1989.

[25] Paul Debevec. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '98, pages 189–198, New York, NY, USA, 1998. ACM.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Jouranl of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[27] Aristeidis Diplaros, Theo Gevers, and Ioannis Patras. Combining color and shape information for illumination-viewpoint invariant object recognition. *IEEE Transactions on Image Processing*, 15:1–11, 2006.

[28] M. Falcone and M. Sagona. An algorithm for the global solution of the shape-from-shading model. In *Image Analysis and Processing*, volume 1310 of *LNCS*, pages 596–603. 1997.

[29] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[30] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. *Int. J. Comput. Vision*, 70:41–54, October 2006.

[31] Drew M.S. Finlayson, G.D and C.Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 79(1):13–30, 2009.

[32] G.D. Finlayson, M.S. Drew, and C. Lu. Intrinsic images by entropy minimization. In *ECCV*, 2004.

[33] G.D. Finlayson, S.D. Hordley, C. Lu, and M.S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006.

[34] R.A. Fisher. Dispersion on a sphere. *Proc. Royal Soc. London*, 217:295–305, 1953.

[35] Robert T. Frankot, Rama Chellappa, and Senior Member. A Method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:439–451, 1988.

[36] Daniel Freedman and Petros Drineas. Energy minimization via graph cuts: Settling what is possible. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:939–946, 2005.

[37] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40:25–47, 2000. 10.1023/A:1026501619075.

[38] Brian V. Funt, Mark S. Drew, and Michael Brockington. Recovering shading from color images. In *ECCV-92: Second European Conference on Computer Vision*, pages 124–132. Springer-Verlag, 1991.

[39] S. Geman and D. Geman. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, pages 452–472. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[40] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

[41] Theo Gevers and Arnold W. M. Smeulders. Color based object recognition. *PR*, 32:453–464, 1999.

[42] Theo Gevers, Joost Van, De Weijer, and Harro Stokman. Color feature detection, 2006.

[43] Amir Globerson and Tommi Jaakkola. Fixing max-product: Convergent message passing algorithms for map lp-relaxations. In *NIPS*, 2007.

[44] Ben Glocker, Nikos Paragios, Nikos Komodakis, Georgios Tziritas, and Nassir Navab. Optical flow estimation with uncertainties through dynamic mrfs. In *CVPR*, 2008.

[45] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006.

[46] Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, pages 2335–2342, 2009.

[47] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR*, pages 2033–2040, 2011.

[48] Alfred Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69:331–371, 1910.

[49] John Haddon and David Forsyth. Shape representations from shading primitives. In *Fifth European Conference on Computer Vision*, pages 415–431, 1997.

[50] John Haddon and David Forsyth. Shading primitives: Finding folds and shallow grooves. In *Proc. Int. Conf. on Computer Vision*, pages 236–241, 1998.

[51] P. L. Hammer, P. Hansen, and B. Simeone. Roof duality, complementation, and persistency in quadratic 0-1 optimization. *Mathematical Programming*, 28:121–155, 1984.

[52] P. L. Hammer, P. Hansen, and B. Simeone. Roof duality, complementation and persistency in quadratic 0-1 optimization. *Mathematical Programming*, 28:121–155, 1984.

[53] P. L. Hammer, P. Hansen, and B. Simeone. Roof duality, complementation and persistency in quadratic 0-1 optimization. *Mathematical Programming*, 28:121–155, 1984.

[54] Feng Han and Song-Chun Zhu. A two-level generative model for cloth representation and shape from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1230–1243, July 2007.

[55] Kenji Hara, Ko Nishino, and Katsushi Ikeuchi. Light source position and reflectance estimation from a single view without the distant illumination assumption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):493–505, 2005.

[56] Kenji Hara, Ko Nishino, and Katsushi Ikeuchi. Mixture of spherical distributions for single-view relighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):25–35, 2008.

[57] K.M. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. In *CVPR*, 2009.

[58] Guenter Hetzel, Bastian Leibe, Paul Levi, and Bernt Schiele. 3d object recognition from range images using local feature histograms. In *Proceedings of CVPR 2001*, pages 394–399, 2001.

[59] B. K.P. Horn. Shape from shading: a method for obtaining the shape of a smooth opaque object from one view. Technical report, Cambridge, MA, USA, 1970.

[60] Berthold K. P. Horn. *Obtaining shape from shading information.*

[61] Berthold K. P. Horn. *Shape from shading.* MIT Press, Cambridge, MA, USA, 1989.

[62] Berthold K. P. Horn. Height and gradient from shading. *Int. J. Comput. Vision*, 5:37–75, September 1990.

[63] D.R. Hougen and N. Ahuja. Estimation of the light source distribution and its use in integrated shape recovery from stereo and shading. In *Fourth International Conference on Computer Vision*, pages 148 – 155, 1993.

[64] Katsushi Ikeuchi and Berthold K. P. Horn. Numerical shape from shading and occluding boundaries. pages 245–299, 1989.

[65] H. Ishikawa. Higher-order clique reduction in binary graph cut. In *CVPR*, 2009.

[66] Hiroshi Ishikawa. Exact optimization for markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1333–1336, 2003.

[67] Hiroshi Ishikawa. Higher-order gradient descent by fusion-move graph cut. In *ICCV*, pages 568–574, 2009.

[68] Hiroshi Ishikawa. Transformation of general binary mrf minimization to the first-order case. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1234–1249, 2011.

[69] M. I. Jordan. *Learning in Graphical Models (Adaptive Computation and Machine Learning)*. MIT Press, 2007.

[70] Olivier Juan and Yuri Boykov. Active graph cuts. In *CVPR*, pages 1023–1029, 2006.

[71] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH ASIA)*, 30(6), 2011.

[72] C.-Y. Kim, A. P. Petrov, H.-K. Choh, Y.-S. Seo, and I.-S. Kweon. Illuminant direction and shape of a bump. *Journal of the Optical Society of America A*, 15:2341–2350, September 1998.

[73] T. Kim and K.S. Hong. A practical approach for estimating illumination distribution from shadows using a single image. *IJIST*, 15(2):143–154, 2005.

[74] Ron Kimmel and Alfred M. Bruckstein. Tracking level sets by level sets: A method for solving the shape from shading problem. *Computer Vision and Image Understanding*, 62(1):47–58, 1995.

[75] Ron Kimmel, Kaleem Siddiqi, Benjamin B. Kimia, and Alfred M. Bruckstein. Shape from shading: Level set propagation and viscosity solutions. *International Journal of Computer Vision*, 16(2):107–133, 1995.

[76] Jan J. Koenderink and Sylvia C. Pont. Irradiation direction from texture. *J. of the Optical Society of America*, 20:2003, 2003.

[77] P. Kohli, M.P. Kumar, and P.H.S. Torr. P3 and beyond: Solving energies with higher order cliques. pages 1–8, 2007.

[78] P. Kohli, M.P. Kumar, and P.H.S. Torr. P3 and beyond: Move making algorithms for solving higher order functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1645–1656, September 2009.

[79] Pushmeet Kohli and M. Pawan Kumar. Energy minimization for linear envelope mrfs. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1863–1870, 2010.

[80] Pushmeet Kohli, Alexander Shekhovtsov, Carsten Rother, Vladimir Kolmogorov, and Philip Torr. On partial optimality in multi-label MRFs. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 480–487, New York, NY, USA, 2008. ACM.

[81] Pushmeet Kohli and Philip H. S. Torr. Efficiently solving dynamic markov random fields using graph cuts. *Computer Vision, IEEE International Conference on*, 2:922–929, 2005.

[82] Pushmeet Kohli and Philip H. S. Torr. Dynamic graph cuts for efficient inference in markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:2079–2088, 2007.

[83] Pushmeet Kohli and Philip H. S. Torr. Measuring uncertainty in graph cut solutions. *Comput. Vis. Image Underst.*, 112:30–38, October 2008.

[84] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[85] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1568–1583, October 2006.

[86] Vladimir Kolmogorov and Carsten Rother. Minimizing nonsubmodular functions with graph cuts-a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1274–1279, 2007.

[87] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:65–81, 2004.

[88] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *CVPR*, 2009.

[89] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. Mrf optimization via dual decomposition: Message-passing revisited. In *ICCV*, 2007.

[90] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. Mrf optimization via dual decomposition: Message-passing revisited. In *ICCV*, 2007.

[91] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:531–552, 2011.

[92] Nikos Komodakis and Georgios Tziritas. Approximate labeling via graph-cuts based on linear programming. In *Pattern Analysis and Machine Intelligence*, page 2007, 2007.

[93] Nikos Komodakis, Georgios Tziritas, and Nikos Paragios. Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies. *Comput. Vis. Image Underst.*, 112:14–29, October 2008.

[94] Nikos Komodakis, Georgios Tziritas, and Nikos Paragios. Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies. *Computer Vision and Image Understanding*, 112(1):14–29, 2008.

[95] V. K. Koval and M. I. Schlesinger. Two-dimensional programming in image analysis problems. *Automatics and Telemechanics*, 8:149–168, 1976.

[96] M. Pawan Kumar, Vladimir Kolmogorov, and Philip H. S. Torr. An analysis of convex relaxations for map estimation of discrete mrfs. *Journal of Machine Learning Research*, 10:71–106, 2009.

[97] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Estimating natural illumination from a single outdoor image. In *ICCV*, 2009.

[98] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *European Conference on Computer Vision*, 2010.

[99] Xiangyang Lan, Stefan Roth, Daniel P. Huttenlocher, and Michael J. Black. Efficient belief propagation with learned higher-order markov random fields. In *ECCV (2)'06*, pages 269–282, 2006.

[100] S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[101] Yvan G. Leclerc and Aaron F. Bobick. The direct computation of height from shading. In *Conference on Computer Vision and Pattern Recognition*, pages 552–558, 1991.

[102] Chia-Hoang Lee and Azriel Rosenfeld. Improved methods of estimating shape from shading using the light source coordinate system. *Artif. Intell.*, 26(2):125–143, May 1985.

[103] K. M. Lee and C. C. J. Kuo. Shape from shading with a linear triangular element surface model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(8):815–822, 1993.

[104] Kyoung Mu Lee and C.-C. Jay Kuo. Shape from shading with a generalized reflectance map model. *Comput. Vis. Image Underst.*, 67:143–160, August 1997.

[105] Tai Sing Lee, Tom Stepleton, Brian Potetz, and Jason Samonds. *Neural Coding of Scene Statistics for Surface and Object Inference*, pages 451–474. Cambridge University Press, 2009.

[106] Sidney R. Lehky and Terrence J. Sejnowski. Network model of shape-from-shading: neural function arises from both receptive and projective fields. *Nature*, 333:452–454, June 1988.

[107] V. Lempitsky, C. Rother, , and A. Blake. Logcut - efficient graph cut optimization for markov random fields. In *ICCV*, 2007.

[108] Victor Lempitsky, Carsten Rother, Stefan Roth, and Andrew Blake. Fusion moves for markov random field optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1392–1405, 2010.

[109] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, 2008.

[110] F.F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, April 2007.

[111] Yuanzhen Li, Stephen Lin, Hanqing Lu, and Heung-Yeung Shum. Multiple-cue illumination estimation in textured scenes. In *ICCV*, 2003.

[112] Yuanzhen Li, Stephen Lin, Hanqing Lu, and Heung yeung Shum. Multiple-cue illumination estimation in textured scenes. In *IEEE Proc. 9th International Conference on Computer Vision*, pages 1366–1373, 2003.

[113] I. Malik and D. Maydan. Recovering three-dimensional shape from a single image of curved objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:555–566, June 1989.

[114] Stephen R. Marschner and Donald P. Greenberg. Inverse lighting for photography. In *Fifth Color Imaging Conference*, pages 262–265, 1997.

[115] Xue Mei, Haibin Ling, and David W. Jacobs. Illumination recovery from image with cast shadows via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[116] E. Mingolla and J. T. Todd. Perception of solid shape from shading. *Biological Cybernetics*, 53:137–151, 1986. 10.1007/BF00342882.

[117] Daisuke Miyazaki, Robby T. Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *Proc. IEEE Intl Conf. Computer Vision*, pages 982–987, 2003.

[118] J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, December 2007.

[119] Ren Ng. All-frequency shadows using non-linear wavelet lighting approximation. *ACM Transactions on Graphics*, 22:376–381, 2003.

[120] Peter Nillius and Jan olof Eklundh. Automatic estimation of the projected light source direction. In *CVPR*, pages 1076–1083, 2001.

[121] Ko Nishino. Directional statistics brdf model. In *ICCV*, pages 476–483, 2009.

[122] Takahiro Okabe, Imari Sato, and Yoichi Sato. Spherical harmonics vs. haar wavelets: Basis for recovering illumination from cast shadows. In *Shadows, Proc. Conf. Computer Vision and Pattern Recognition*, pages 50–57, 2004.

[123] Takayuki Okatani and Koichiro Deguchi. Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. *Comput. Vis. Image Underst.*, 66:119–131, May 1997.

[124] J. Oliensis. Shape from shading as a partially well-constrained problem. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 559–564, 1991.

[125] J. Oliensis and P. Dupuis. A global algorithm for shape from shading. In *ICCV'93*, pages 692–701, 1993.

[126] John Oliensis. Uniqueness in shape from shading. *International Journal of Computer Vision*, 6(2):75–104, 1991.

[127] John Oliensis and Paul Dupuis. Shape recovery. chapter Direct method for reconstructing shape from shading, pages 17–28. Jones and Bartlett Publishers, Inc., , USA, 1992.

[128] A. Panagopoulos, D. Samaras, and N. Paragios. Robust shadow and illumination estimation using a mixture model. In *CVPR*, 2009.

[129] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios. Illumination estimation and cast shadow detection through a higher-order graphical model. In *CVPR*, 2011.

[130] Tai pang Wu and Chi keung Tang. A bayesian approach for shadow extraction from a single image. In *Proc. of the tenth International Conference on Computer Vision*, pages 480–487, 2005.

[131] J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the American Association of Artificial Intelligence National Conference on AI*, pages 133–136, Pittsburgh, PA, 1982.

[132] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[133] M.A. Penna. A shape from shading analysis for a single perspective image of a polyhedron. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:545–554, 1989.

[134] A. Pentland. Shape information from shading: A theory about human perception. In *Computer Vision., Second International Conference on*, pages 404–413, 1988.

[135] Alex P. Pentland. Local shading analysis. pages 443–487, 1989.

[136] Alex P. Pentland. Shape from shading. chapter Local shading analysis, pages 443–487. MIT Press, Cambridge, MA, USA, 1989.

[137] A.P. Pentland. Finding the illuminant direction. *Journal of the Optical Society of America*, 72:448, April 1982.

[138] Brian Potetz. Efficient belief propagation for vision using linear constraint nodes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07)*. IEEE Computer Society, Minneapolis, MN, USA, 2007.

[139] Brian Potetz and Tai Sing Lee. Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. *Journal of the Optical Society of America A*, 20(7):1292–1303, Jul 2003.

[140] Brian Potetz and Tai Sing Lee. Scaling laws in natural scenes and the inference of 3D shape. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1089–1096. MIT Press, Cambridge, MA, 2006.

[141] Brian Potetz and Tai Sing Lee. Efficient belief propagation for higher order cliques using linear constraint nodes. *Computer Vision and Image Understanding*, 112(1):39–54, Oct 2008.

[142] Mark W. Powell, Student Member, Sudeep Sarkar, Dmitry Goldgof, and Senior Member. A simple strategy for calibrating the geometry of light sources. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1022–1027, 2001.

[143] E. Prados and O. Faugeras. "perspective shape from shading" and viscosity solutions. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 826–, Washington, DC, USA, 2003. IEEE Computer Society.

[144] Emmanuel Prados. A unifying and rigorous shape from shading method adapted to realistic data and applications. In *Journal of Mathematical Imaging and Vision*, pages 307–328, 2006.

[145] Emmanuel Prados and Olivier Faugeras. A generic and provably convergent shape-from-shading method for orthographic and pinhole cameras, int. *J. Computer Vision*, 65:97–125, 2005.

[146] Emmanuel Prados and Olivier Faugeras. Shape from shading: a well-posed problem ? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, California*, volume II, pages 870–877. IEEE, Jun 2005.

[147] Hossein Ragheb and Edwin R. Hancock. A probabilistic framework for specular shape-from-shading. *Pattern Recognition*, 36(2):407–427, 2003.

[148] Ramachandran and S. Vilayanur. Perceiving Shape from Shading. *Scientific American*, 259:76–83, August 1988.

[149] Srikumar Ramalingam, Pushmeet Kohli, Karteek Alahari, and Philip H. S. Torr. Exact inference in multi-label crfs with higher order cliques. In *CVPR*, 2008.

[150] R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object. *Journal of the Optical Society of America A*, 18:2448–2459, October 2001.

[151] Ravi Ramamoorthi. Analytic pca construction for theoretical analysis of lighting variability in images of a lambertian object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1322–1333, 2002.

[152] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for reflection. *ACM TRANSACTIONS ON GRAPHICS*, 23:1004–1042, 2004.

[153] Ravi Ramamoorthi, Melissa Koudelka, and Peter Belhumeur. A fourier theory for cast shadows. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 2005, 2004.

[154] I. Rosenberg. Reduction of bivalent maximization to the quadratic case. *Cahiers du Centre d'Etudes de Recherche Operationnelle*, 1975.

[155] Stefan Roth and Michael Black. Fields of experts. *International Journal of Computer Vision*, 82:205–229, 2009. 10.1007/s11263-008-0197-6.

[156] C. Rother, P. Kohli, Wei Feng, and Jiaya Jia. Minimizing sparse higher order energy functions of discrete variables. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1382–1389, 2009.

[157] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23:309–314, 2004.

[158] Carsten Rother, Vladimir Kolmogorov, Victor Lempitsky, and Martin Szummer. Optimizing binary mrfs via extended roof duality. In *CVPR*, 2007.

[159] Elisabeth Rouy and Agnès Tourin. A Viscosity Solutions Approach to Shape-From-Shading. *SIAM Journal on Numerical Analysis*, 29(3):867–884, 1992.

[160] Ruslan Salakhutdinov. Learning in markov random fields using tempered transitions. In *In Advances in Neural Information Processing Systems*, page 2010.

[161] Elena Salvador, Andrea Cavallaro, and Touradj Ebrahimi. Cast shadow segmentation using invariant color features. *CVIU*, 95(2):238–259, 2004.

[162] Dimitrios Samaras and Dimitris Metaxas. Coupled lighting direction and shape estimation from single images. *Computer Vision, IEEE International Conference on*, 2:868, 1999.

[163] Dimitrios Samaras, Dimitris Metaxas, P Ascalfua, and Yvan G. Leclerc. Variable albedo surface reconstruction from stereo and shape from shading. In *CVPR*, pages 480–487, 2000.

[164] Dimitris Samaras and Dimitris Metaxas. Incorporating illumination constraints in deformable models for shape from shading and light direction estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:247–264, February 2003.

[165] Imari Sato, Yoichi Sato, and Katsushi Ikeuchi. Illumination distribution from shadows. In *CVPR*, 1999.

[166] Imari Sato, Yoichi Sato, and Katsushi Ikeuchi. Illumination from shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):290–300, 2003.

[167] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2000.

[168] Sameer Shirdhonkar and David W. Jacobs. Non-negative lighting and specular object recognition. In *ICCV 05*, pages 1323–1330, 2005.

[169] Yael Shor and Dani Lischinski. The shadow meets the mask: Pyramid-based shadow removal. *Computer Graphics Forum*, 27(2):577–586, apr 2008.

[170] Richard Szeliski. Fast Shape from Shading. In *ECCV '90: Proceedings of the First European Conference on Computer Vision*, pages 359–368, London, UK, 1990. Springer-Verlag.

[171] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1068–1080, 2008.

[172] Marshall F Tappen, William T Freeman, and Edward H Adelson. Recovering intrinsic images from a single image, 2002.

[173] Marshall F. Tappen, William T. Freeman, and Edward H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1459–1472, 2005.

[174] Marshall F. Tappen, William T. Freeman, and Edward H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1459–1472, September 2005.

[175] Daniel Tarlow, Inmar Givoni, and Richard Zemel. Hop-map: Efficient message passing with high order potentials. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

[176] Shoji Tominaga and Norihiro Tanaka. Estimating reflection parameters from a single color image. *IEEE Comput. Graph. Appl.*, 20:58–66, September 2000.

[177] P. S. Tsai and M. Shah. Shape From Shading Using Linear Approximation. *IVC*, 12(8):487–498, October 1994.

[178] Greg Turk and Marc Levoy. Zippered polygon meshes from range images. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, SIGGRAPH '94, pages 311–318, New York, NY, USA, 1994. ACM.

[179] Joost van de Weijer, Theo Gevers, and Jan-Mark Geusebroek. Edge and corner detection by photometric quasi-invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 2005.

[180] Joost van de Weijer, Theo Gevers, and Jan-Mark Geusebroek. Edge and corner detection by photometric quasi-invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:625–630, April 2005.

[181] Manik Varma and Andrew Zisserman. Estimating illumination direction from textured images. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:179–186, 2004.

[182] A. Varol, A. Shaji, M. Salzmann, and P. Fua. Monocular 3d reconstruction of locally textured surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2011.

[183] O. E. Vega and Y. H. Yang. Default shape theory: With application to the computation of the direction of the light source. *CVGIP: Image Understanding*, 60(3):285 – 299, 1994.

[184] O.E. Vega and Y.H. Yang. Shading logic: A heuristic approach to recover shape from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:592–597, 1993.

[185] Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky. Map estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. *IEEE Transactions on Information Theory*, 51:3697–3717, 2005.

[186] Chaohui Wang, Martin De la Gorce, and Nikos Paragios. Segmentation, ordering and multi-object tracking using graphical models. In *ICCV*, 2009.

[187] Yang Wang and Dimitris Samaras. Estimation of multiple directional light sources for synthesis of augmented reality images. *Graphical Models (Special Issue on Pacific Graphics)*, 65(4):185–205, 2003.

[188] Gregory J. Ward. Measuring and modeling anisotropic reflection. *SIGGRAPH Comput. Graph.*, 26:265–272, July 1992.

[189] Guo-Qin Wei and G. Hirzinger. Learning shape from shading by a multilayer network. *Neural Networks, IEEE Transactions on*, 7(4):985–995, 1996.

[190] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12:1–41, 2000.

[191] Yair Weiss and William T. Freeman. Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology. *Neural Computation*, 13(10):2173–2200, October 2001.

[192] Philip L. Worthington and Edwin R. Hancock. New constraints on data-closeness and needle map consistency for shape-from-shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:1250–1267, December 1999.

[193] Y. Yang and A.L. Yuille. Source from shading. In *CVPR91*, pages 534–539, 1991.

[194] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Generalized belief propagation. In *NIPS 13*, pages 689–695. MIT Press, 2000.

[195] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Analysis of shape from shading techniques. In *CVPR*, pages 377–384, 1994.

[196] Yufei Zhang and Yee-Hong Yang. Illuminant direction determination for multiple light sources. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:1269, 2000.

[197] Yufei Zhang and Yee-Hong Yang. Multiple illuminant direction detection with application to image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:915–920, August 2001.

[198] Qinfen Zheng and Rama Chellappa. Estimation of illuminant direction, albedo, and shape from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):680–702, 1991.

[199] W. Zhou and C. Kambhamettu. A unified framework for scene illuminant estimation. *IVC*, 26(3):415–429, 2008.

[200] J. Zhu, K. G. G. Samuel, S. Masood, and M. F. Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, 2010.